CHOOSING PRIOR HYPERPARAMETERS: WITH APPLICATIONS TO TIME-VARYING PARAMETER MODELS

Pooyan Amir-Ahmadi, Christian Matthes and Mu-Chun Wang

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: Priors, Bayesian inference, Bayesian VAR, Time variation.

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. While this feature is present to some degree not only in Vector Autoregressions (VARs, Sims (1980)), but also in multivariate time series models that try to economize on the number of parameters such as factor models (see for example Stock & Watson (2002)) or Dynamic Stochastic General Equilibrium (DSGE) models (Smets & Wouters (2007)), it is maybe most evident in VARs that feature time-varying parameters and stochastic volatility.¹

Affiliations: University of Illinois at Urbana-Champaign (Amir-Ahmadi, pooyan@illinois.edu), Federal Reserve Bank of Richmond (Matthes, christian.matthes@gmail.com), and University of Hamburg (Wang, Mu-Chun.Wang@wiso.uni-hamburg.de).

¹We would like to thank Luca Benati, Fabio Canova, Minsu Chan, 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.

¹This class of models has been introduced by the work of Cogley & Sargent (2005) and Primiceri (2005), building on earlier contributions such as Doan et al. (1984), Stock & Watson (1996) and Cooley & Prescott (1976).
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)).

Our main focus in terms of applications in this paper is on the aforementioned VARs with time-varying parameters and stochastic volatility. The work of Cogley & Sargent (2005) and Primiceri (2005) has established a de-facto standard on how to set the priors for these models using a training sample and relatively few hyperparameters. In this paper, we show in a Monte Carlo exercise that using our approach, instead of relying on standard values for the hyperparameters, can drastically improve the estimation of the parameter and volatility paths as well as the forecasting performance of these models. We also show that our approach can lead to substantially different conclusions in applications with real data - we study the joint dynamics of US and UK inflation for a monthly sample covering the last century.

The Gibbs sampler is already the standard approach to estimate models in this class and thus our approach fits naturally into the estimation approach used for these models.\footnote{The Gibbs sampler can be viewed as a special case of the Metropolis-Hastings algorithm (see again Gelman et al. (2013))}

\footnote{Examples of papers that follow this structure are, among many others, Canova & Gambetti (2009), Clark & Terry (2010), Baumeister & Peersman (2013), Benati & Lubik (2014), Kliem et al. (2016), Koop et al. (2011), and D’Agostino & Surico (2011).}

\footnote{Even if the model of interest has previously been estimated with another posterior sampler, our approach can be used in that case - the previous posterior sampler then forms one block of a multiple-block algorithm,
Our approach allows researchers to estimate hyperparameters for the time variation in all parameters (including the volatilities of the residuals) without restricting those parameters to possibly take on only a small finite number of values.

The importance of hyperparameters in this class of models 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’s approach to estimating the prior hyperparameters is computationally very 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 used should be employed for other applications.

Since Primiceri (2005), most applications of models in this class use his estimated values for the hyperparameters or other fixed values on an ad hoc basis. Furthermore, these papers do not take into account that uncertainty about these hyperparameters could influence later inference, whereas our approach automatically takes this uncertainty into account.

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.

While this paper focuses on presenting our approach using a time-varying parameter VAR with stochastic volatility, the extension to other time-varying parameter models such as factor-augmented VARs with time-varying parameters and stochastic volatility (see Ellis et al. (2014) and Baumeister et al. (2013), for example) or factor models with time-varying parameters and stochastic volatility is straightforward because of the modular structure of the Gibbs sampler. Our approach will also be of interest to researchers trying to estimate multivariate extensions of the univariate stochastic volatility model presented in Kim et al. (1998). As will become evident in our discussion of the models we study, this class of models allows for the stochastic increments in the stochastic volatility processes for different variables to be correlated, whereas in the finance literature on multivariate stochastic volatility, those increments are often assumed to be independent (Chib et al. (2006)). Our approach can also

---

5To be more specific, Primiceri (2005) uses a reversible jump MCMC algorithm to estimate the hyperparameters. To get proposal densities, he in a first step has to estimate his time-varying parameter model conditional on each possible combination of hyperparameters.

6This echoes the results in Reusens & Croux (2015), who carry out an extensive Monte Carlo study of prior sensitivity using a VAR with time-varying parameters but no stochastic volatility.

7In the appendix, we show how to augment the standard Gibbs sampler for fixed coefficient VARs to include the estimation of prior hyperparameters.
be used for DSGE models with stochastic volatility as introduced by Justiniano & Primiceri
(2008). Incorporating such stochastic volatility in DSGE models substantially improves the
forecasting performance, as highlighted by Diebold et al. (In Press).

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.

Lubik et al. (2014) calibrate the hyperparameters using a prior predictive approach: They
simulate paths for parameters and volatilities from the prior and choose fixed values for
the hyperparameters so that a pre-specified set of moments of the simulated parameters
and volatilities approximate as close as possible the corresponding moments obtained by
estimating fixed coefficients VARs on a set of rolling training samples.

Our paper is more generally related to the literature on choosing prior hyperparameters in
Bayesian inference. Giannone et al. (2013) 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. 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. (2013) 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. (2013) can thus not be easily
extended to time-varying parameter models. As highlighted by Giannone et al. (2013), their
first approach (which maximizes the marginal likelihood) is an empirical Bayes approach,

8In the second case they then propose to draw the other VAR parameters conditional on the hyperpara-

meters. 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.
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. (2013), 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. (2013), 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 infeasible 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. Our approach, on the other hand, is not sensitive to the priors for the other parameters in the model.

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 a real data set of historical inflation data for the US and the UK in section 5.

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 \( \kappa \). 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. 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 $\kappa_j$ ($j = 1, \ldots, J$) and the vector of all hyperparameters by $\bar{\kappa} = [\kappa'_1 \ k'_2 \ \ldots \ k'_J]'$. 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 assumptions hold (assumption 1 is only necessary if multiple vectors of hyperparameters are present in the model):

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

**Assumption 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 assumptions are restrictive. If assumption 1 is violated, the dependent vectors of hyperparameters just have to be grouped into one larger vector of hyperparameters. The modifications for the algorithm in this case are straightforward. Violations of the

---

9 The appendix shows how to use our approach with a popular specification for fixed coefficient VARs.

10 We later spell out these assumptions in more detail for our VAR model.
second assumption 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

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

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

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

By the hierarchical nature of the model (the hyperparameters only enter the prior for $K$), $p(Y|\theta, \kappa, K)$ 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 assumption 2 and the hierarchical nature of the hyperparameter structure (and, if necessary, assumption 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

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

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. In the appendix, we lay out the estimation algorithm for this class of models in detail.

The observable vector $y_t$ is modeled as:

$$y_t = \mu_t + \sum_{j=1}^{L} B_{j,t} y_{t-j} + e_t$$

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 $X_t' \equiv I \otimes (1, y_{t-1}', \ldots, y_{t-L}')$, $b_t \equiv vec (\mu_t B_{1,t} \cdots B'_{L,t})$ and rewrite (3.1) in the following state space form\textsuperscript{12}:

$$y_t = X_t' b_t + e_t$$

$$b_t = b_{t-1} + \omega_{b,t}$$

The observation equation (3.2) is a more compact expression for (3.1). The state equation (3.3) describes the law of motion for the intercepts and autoregressive matrices. The covariance matrix of the innovations in equation (3.2) is modeled following Primiceri (2005):

$$e_t = A_t^{-1} \Sigma_t^{5} \varepsilon_t$$

The covariance state $A_t$ is a lower triangular matrix with ones on the main diagonal and representative non fixed element $a^i_t$. $\Sigma_t$ is a diagonal matrix with representative non fixed element $\sigma^j_t$. The dynamics of the non fixed elements of $A_t$ and $\Sigma_t^{5}$ are given by:

$$a^i_t = a^i_{t-1} + \omega^i_{a,t}$$

$$\log \sigma^j_t = \log \sigma^j_{t-1} + \omega^j_{h,t}$$

To conclude the description of our model, we need to make distributional assumptions 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 $\Sigma_t^{5}$ 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

\textsuperscript{12} I denotes a identity matrix of conformable size and 1 denotes a vector of ones of conformable size.
the following covariance matrix, which we, following Primiceri, restrict as follows:

\[
\begin{pmatrix}
\varepsilon_t \\
\omega_{b,t} \\
\omega_{a,t} \\
\omega_{h,t}
\end{pmatrix}
\begin{bmatrix}
I & 0 & 0 & 0 \\
0 & \Omega_b & 0 & 0 \\
0 & 0 & \Omega_a & 0 \\
0 & 0 & 0 & \Omega_h
\end{bmatrix}
\]

\(\Omega_a\) 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. We repeat, for expository purposes, the derivation of the Metropolis-Hastings acceptance probability from the previous section as applied to this specific model.

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

\[
\begin{align*}
\Omega_b & \sim p_{\Omega_b}(\kappa_{\Omega_b}) \\
\Omega_h & \sim p_{\Omega_h}(\kappa_{\Omega_h}) \\
\Omega_{a,j} & \sim p_{\Omega_{a,j}}(\kappa_{\Omega_a}) \quad \forall j = 1, \ldots, J
\end{align*}
\]

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 (3.8) to (3.10) 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. Bayesian inference lends itself naturally to the estimation of hierarchical models because they can be analyzed with a Gibbs sampler or the multiple-block Metropolis-Hastings algorithm. This ease of estimation and the flexibility of hierarchical models explains their popularity (see for example Gelman et al. (2013)).

We now restate assumptions 1 and 2 for the specific model at hand:
Assumption 3  The different vectors of hyperparameters in a TVP-VAR are *a priori* independent of each other:

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

Assumption 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 3.8 through 3.10 in this model) are *a priori* independent of that specific hyperparameter (e.g. \( \Omega_h \) and \( \Omega_{a,j} \) \( \forall j = 1, \ldots, J \) 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:

\[
\alpha^i = \min \left( 1, \frac{p(\theta, \kappa_{\Omega_b}^{\text{prop}}, \Omega_b, \kappa_{\Omega_a}, \{\Omega_{a,j}\}, \kappa_{\Omega_h}, \Omega_h | y^T) q(\kappa_{\Omega_b}^{\text{prop}} | \kappa_{\Omega_b}^{i-1})}{p(\theta, \kappa_{\Omega_b}^{i-1}, \Omega_b, \kappa_{\Omega_a}, \{\Omega_{a,j}\}, \kappa_{\Omega_h}, \Omega_h | y^T) q(\kappa_{\Omega_b}^{i-1} | \kappa_{\Omega_b}^{\text{prop}})} \right)
\]

where a superscript \( \text{prop} \) 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.

For each vector of hyperparameters, we need to calculate the following posterior:

\[
p(\theta, \kappa_{\Omega_b}, \Omega_b, \kappa_{\Omega_a}, \{\Omega_{a,j}\}, \kappa_{\Omega_h}, \Omega_h | y^T) \propto p(y^T | \theta, \kappa_{\Omega_b}, \Omega_b, \kappa_{\Omega_a}, \{\Omega_{a,j}\}, \kappa_{\Omega_h}, \Omega_h) \\
\times p(\theta, \kappa_{\Omega_a}, \{\Omega_{a,j}\}, \kappa_{\Omega_h}, \Omega_h | \kappa_{\Omega_b}, \Omega_b) \\
\times p(\Omega_b | \kappa_{\Omega_b}) p(\kappa_{\Omega_b})
\]

Two observations are key to simplifying this formula: First, once we condition on \( \theta \) and the various \( \Omega \) matrices, the conditional density of \( y^T \) is independent of \( \kappa_{\Omega_b} \) and thus cancels out in the calculation of the Metropolis-Hastings acceptance probability because it appears in both the denominator and the numerator. This can be seen by studying equation (3.1) and noting that \( \theta \) contains the sequence of \( \{\mu_t, \{B_{j,t}\}_{j=1}^L, A_t, \Sigma_t\}_{t=1}^5 \). Once we know \( \theta \) and \( \Omega_b \),
CHOOSING PRIOR HYPERPARAMETERS

$\kappa_{\Omega_b}$ is not needed for the calculation of this density. The second point worth noting is that conditional on $\Omega_b$, $\kappa_{\Omega_b}$ does not carry any information about any of the other parameters. This is due to the hierarchical nature of the model and the two assumptions made before. Thus we get that

$$p(\theta, \kappa_{\Omega_a}, \{\Omega_{a_j}\}, \kappa_{\Omega_b}, \Omega_b) = p(\theta, \kappa_{\Omega_a}, \{\Omega_{a_j}\}, \kappa_{\Omega_b}, \Omega_b|\Omega_b)$$

As a result, we find that $p(\theta, \{\kappa_{\Omega_a}\}, \{\Omega_{a_j}\}, \kappa_{\Omega_h}, \Omega_h|\Omega_b)$ also cancels out in the acceptance probability since it is not a function of $\kappa_{\Omega_b}$. The acceptance probability then simplifies to

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

$p(\Omega_b|\kappa_{\Omega_b})$ 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:

$$\Omega_b \sim IW\left(\kappa_{\Omega_b}^2 \nu_{\Omega_b} V_{\Omega_b}, \nu_{\Omega_b}\right)$$

(3.15)

(3.16)

(3.17)

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 $\kappa_X^4$.\(^{13}\)

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

\(^{13}\)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.
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

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

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 $\kappa_x$ 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 $\Omega_{a,j}$:

$$\Omega_b \sim IW(diag(\kappa_{\Omega_b})\nu_{\Omega_b} V_{\Omega_b} diag(\kappa_{\Omega_b}), \nu_{\Omega_b})$$

$$\Omega_h \sim IW(diag(\kappa_{\Omega_h})\nu_{\Omega_h} V_{\Omega_h} diag(\kappa_{\Omega_h}), \nu_{\Omega_h})$$

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

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 $\tilde{\kappa}_x$) to $\kappa_x$.

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 (3.19)) 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 $\kappa_{\Omega_b}$, 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

\footnote{The changes are the same for all blocks. We describe it here for the case of $\Omega_b$.}
density needs to be updated as described in equation (3.19).

Various choices for the priors of the hyperparameters are possible. While a uniform prior might seem attractive at first sight, it has two problems: (i) 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) and (ii) for some time-varying parameter models such as local level models, pile-up problems at a hyperparameter implying a signal to noise ratio of 0 can occur (see Shephard (1993)). The second point means that the likelihood function can have a peak at a hyperparameter implying no time variation when in fact the true data-generating process features time variation. To circumvent this problem, we use inverse-gamma priors as a benchmark for our hyperparameters, but display results for various priors in our Monte Carlo study.

4. MONTE CARLO STUDY

We use a univariate AR(1) process in a simulation study to demonstrate the properties of our approach. The issue of choosing hyperparameters in time-varying parameter models arises in univariate models as well as in multivariate models, so we choose a univariate model for simplicity. We estimate AR(1) models with time-varying intercept and AR parameter as well as stochastic volatility:

\[
y_t = \mu_t + \phi_t y_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma_t^2)
\]

\[
\begin{pmatrix}
\mu_t \\
\phi_t
\end{pmatrix}
= \begin{pmatrix}
\mu_{t-1} \\
\phi_{t-1}
\end{pmatrix}
+ \begin{pmatrix}
e_{1,t} \\
e_{2,t}
\end{pmatrix}, \quad e_t \overset{iid}{\sim} N(0, \Omega_b)
\]

\[
h_t = h_{t-1} + u_t, \quad u_t \overset{iid}{\sim} N(0, \Omega_h)
\]

Let us define the \(h_t \equiv \log \sigma_t\) and \(b_t \equiv (\mu_t, \phi_t)'\) for this section. To assess how well our approach works relative to a model with fixed hyperparameters, we study the root mean squared error for the parameter path. To be more specific, the statistic we use is

\[
RMSE = \sqrt{\frac{1}{N_{sim}} \sum_{j=1}^{N_{sim}} \frac{1}{T} \sum_{t=1}^{T} \left( \hat{\theta}_{t,j} - \theta_{t,j} \right)^2}
\]

where \(\hat{\theta}_{t,j}\) denotes the estimated posterior median of the coefficient at time \(t\) in Monte Carlo sample \(j\). \(\theta_{t,j}\) denotes the corresponding true value. \(T\) is the sample size for each simulation,
which we set to 350. 40 periods of these 350 are used as a training sample for the prior. \( N_{sim} \) is the total number of Monte Carlo repetitions, which we set to 100. The relative \( RMSE \) is computed relative to the fixed hyperparameter specification (we use the values from Primiceri (2005) as our benchmark). In addition, we also want to investigate whether any possible gains

\[
RMSFE_h = \sqrt{\frac{1}{N_{sim}} \sum_{j=1}^{N_{sim}} (\hat{y}_{t+h,j} - y_{t+h,j})^2}
\]

where \( \hat{y}_{t+h,j} \) denotes the median \( h \)-step ahead forecast of Monte Carlo sample \( j \). \( y_{t+h,j} \) denotes the corresponding true value. For both root mean squared error criteria we reports values relative to the case with fixed hyperparameters.

We use various data-generating processes to highlight that these models can, when hyperparameters are estimated, successfully recover the path of parameters even when these parameters do not follow a random walk and the model is thus misspecified. First, we use a data-generating process that is correctly specified, except that we impose that \( \phi_t < 1 \forall t \) so that the resulting time series resemble the time series we actually observe in real-world
applications.\textsuperscript{15} We initialize the intercept and the AR coefficient at 0 and the stochastic volatility at 0.1. We set $\Omega_b$ to be a diagonal matrix with diagonal elements 0.0001 and $\Omega_h$ to 0.001. Table 4.1 shows the results for this exercise. We can see that in terms of parameters, our approach substantially outperforms the fixed hyperparameter approach.\textsuperscript{16} We show results for various prior distributions for the hyperparameters (we use the same prior for each hyperparameter within one estimation), adding the half-Cauchy and half-t distributions (as suggested by Gelman (2006)) to our benchmark inverse-Gamma prior. The half-t and half-Cauchy distribution do not, as emphasized by Gelman (2006), make a hyperparameter value close to 0 very unlikely, but can rather be parametrized to be smooth around 0 so they might be preferable in some situations. To be specific, the priors we use are either a inverse-Gamma distribution with a scale parameter of 1/15 and 6 degrees of freedom, a half-Cauchy distribution with a scale parameter of 1/15 or a half-t distribution with a scale parameter of 1/15 and 6 degrees of freedom. Figure 4.1 displays these priors.

\begin{table}[h]
\centering
\caption{Monte Carlo results for $DGP_1$ with random walk evolution of parameters.}
\begin{tabular}{lcccc}
\hline
Parameter & $i\mathcal{G}$ & half-Cauchy & half-t & Fixed \\
$\mu_t$ & 0.7037 & 0.7326 & 0.7519 & 1.0000 \\
$\phi_t$ & 0.6300 & 0.6471 & 0.6557 & 1.0000 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Monte Carlo results for $DGP_1$ with random walk evolution of parameters.}
\begin{tabular}{lcccc}
\hline
Horizons & $i\mathcal{G}$ & half-Cauchy & half-t & Fixed \\
1 & 1.0112 & 0.9981 & 0.9869 & 1.0000 \\
2 & 0.9430 & 0.9384 & 0.9594 & 1.0000 \\
3 & 0.9872 & 0.9649 & 0.9872 & 1.0000 \\
4 & 0.9659 & 0.9546 & 0.9639 & 1.0000 \\
5 & 0.9301 & 0.9563 & 0.9610 & 1.0000 \\
6 & 0.9479 & 0.9737 & 0.9605 & 1.0000 \\
7 & 0.9030 & 0.9415 & 0.9446 & 1.0000 \\
8 & 0.9297 & 0.9238 & 0.9317 & 1.0000 \\
\hline
\end{tabular}
\end{table}

All those priors outperform the fixed hyperparameter case, with the decreases in root mean squared error being between 25 and 35 percent. This increase in performance translates to forecasting performance, as can be seen in table 4.1. We next focus on a data-generating process that is more severely misspecified, namely parameters that evolve according to sine and

\begin{footnotesize}
\textsuperscript{15}If we do not impose this restriction, simulated time series from this class of models can become explosive.

\textsuperscript{16}Our approach and the fixed hyperparameter approach perform similarly for the stochastic volatility, which is why we omit it here.
\end{footnotesize}
cosine functions. We will keep this DGP fixed across the 100 Monte Carlo samples, allowing for easier visual inspection of the performance of the various hyperparameter settings. We set $\mu_t = -\cos(t)$, $\phi_t = 0.4(\sin(t) + 1)$ and $\sigma_t = \sin(t) + 1.5$. Table 4.2 shows that in this case our approach fares even better than with the first DGP. Figure 4.2 plots the true path of the parameters as well as the 5th percentile, the 95th percentile and the median across Monte Carlo samples of the posterior median paths of the parameters for the case of the inverse gamma prior and the fixed hyperparameter case (the estimated paths for the other priors look similar to the inverse-Gamma case). We see that while our approach does capture the true DGP reasonably well, the fixed hyperparameter case wrongly finds that the parameters do not move much over time. It is worth pointing out that the fixed hyperparameter setup often finds little time variation in many parameters in real world applications such as Cogley & Sargent (2005).

A natural question to ask is how much of this advantage is due to the specific values 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 the appendix we show that even setting $k_{0b}$ 10 times larger than in these benchmark Monte Carlo simulations still leaves our approach superior (and the magnitudes of the differences in root mean squared errors still large).
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?\textsuperscript{17} 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. Nonetheless, our approach is capable of finding basically zero time variation in the parameters when there is none, as highlighted in table 4.3. Both in terms of parameter estimates and forecasting ability our approach and the fixed hyperparameter approach are very similar in this case.

The main takeaway from this exercise is not that the hyperparameters in Primiceri (2005) are ‘wrong’ in any sense, but rather that, if a researcher is interested 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 way to take the dependence of the results on the prior hyperparameters into account.

\textsuperscript{17}We use $\mu = 0.5$, $\phi = 0.8$ and $\sigma = 0.1$ for the data-generating process.
5. EMPIRICAL APPLICATION

VARs with time-varying parameters and stochastic volatility have previously been used to analyze long historical time series - see, for example, Sargent & Surico (2011), D’Agostino & Surico (2011), Kliem et al. (2016), and Amir-Ahmadi et al. (2016). Other papers that have used time series models with time-varying parameters and stochastic volatility (not necessarily VARs) include Cogley & Sargent (2015) and Schorfheide 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 long historical samples. With long historical time series that are very different from the time series used by Primiceri (2005), there is little reason to believe a-priori that the hyperparameters chosen by Primiceri should reflect a researcher’s view of the amount of time variation present in the data. To
assess the importance of estimating the prior hyperparameters, a long historical dataset thus seems like a useful laboratory. We study a VAR with monthly year-over-year CPI inflation data for the US and the UK starting in 1915.\footnote{The data is from the Global Financial database. We abstract from measurement error issues here. For a treatment of measurement errors in VAR models with time-varying parameters and stochastic volatility, see Amir-Ahmadi et al. (2016).}

Figure 5.1.— Marginal posterior distributions of the three hyperparameters along with common values used in the literature.

The first 95 months are used as the training sample to initialize the priors. To be consistent with most of the literature, we use 2 lags.\footnote{We use 50000 draws to adaptively pick the parameters of the proposal densities for the hyperparameters and another 50000 draws to actually approximate the posterior.} We use inverse gamma priors for the hyperparameters. For simplicity, we use the same (loose) prior for all three hyperparameters: We pick the parameters for the prior of the hyperparameters such that the prior mode is at 0.05 and the variance is infinite.\footnote{As a robustness check, we will later show results for a prior with finite variance. Our choice of an inverse-gamma prior confronts the possible pile-up problem for the hyperparameters discussed by Stock \&}
the inverse-gamma prior) along with the values used in Primiceri (2005). We can see that there is substantial uncertainty surrounding the hyperparameters, which our approach takes into account. We also see that the posterior modes are substantially different from the values used in Primiceri (2005). Primiceri estimated these values for a specific dataset. Our results hopefully caution the reader to not blindly use fixed values for other applications. As we saw in the case of our simulations, estimating the prior hyperparameters can have a substantial impact on parameter paths and, as a result, statistics derived from the estimated parameter and volatility paths such as impulse responses, forecasts, and variance decompositions. To illustrate this impact, we focus on the 'core inflation' statistic used by Cogley & Sargent (2005). Core inflation is defined here as the implied steady state value of inflation if the parameters of the VAR were fixed at each point in time at their posterior mean.\footnote{We checked that the VAR based on time $t$ posterior mean parameter values is stationary for all $t$ so}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{uk_core_inflation}
\caption{UK core inflation for the US and UK, using either estimated hyperparameters or the values from Primiceri.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{us_core_inflation}
\caption{US core inflation for the US and UK, using either estimated hyperparameters or the values from Primiceri.}
\end{figure}

Watson (1996). We thus reduce the prior probability mass for the hyperparameters very close to zero. As seen in our Monte Carlo study, this does not preclude finding no time variation in parameters if indeed the data-generating process does not feature time-varying parameters. While the pile-up problem described in Stock & Watson (1996) was not an issue in our Monte Carlo exercise, we do suggest researchers use an informative prior to steer clear of any possibility of the pile-up problem appearing in real-world applications.\footnote{We checked that the VAR based on time $t$ posterior mean parameter values is stationary for all $t$ so}
measure, depicted in Figure 5.2, can be interpreted as an estimate of trend inflation. Figure 5.2 shows that core inflation based on estimated hyperparameters, while still substantially smoother than the data, is not as smooth as the estimate using the fixed hyperparameter values of Primiceri (2005). To give one example, core inflation for the UK peaks around 10% (in annual terms) with our approach, whereas it is only 6% for the fixed hyperparameter-based estimates. The differences are especially pronounced during the 1930s and the 1970s (periods where inflation was large in absolute value). The greater amount of time variation in parameters implied by the posterior distribution shown before clearly has an impact on the core inflation statistic.

As a robustness check, we re-estimate our model using a different prior. For simplicity, we only change the prior for $k_{\Omega_b}$. We keep the prior mode fixed at the same value as before but now use parameter values that imply a finite variance, namely .01 (whereas the prior that this calculation makes sense.
variance was infinite before). Figure 5.3 shows that the results are indeed robust to these changes in the prior.

We next re-estimate our VAR, allowing for two different values of $k_{\Omega_b}$ - one for the intercepts and one for all other parameters. We still use one hyperparameter each for $W$ and $\{S_j\}_{j=1}^J$. For the sake of brevity, we focus here on results using our infinite variance specification for all hyperparameters (results using the finite variance prior are similar, just as before). Figure 5.4 plots the marginal posteriors of the hyperparameters. We can see an interesting pattern: The posterior distribution for the $k_{\Omega_b}$ value associated with the intercepts is similar to that obtained for all parameters in the previous section, but the $k_{\Omega_b}$ parameters for all other parameters is an order of magnitude larger, showing that substantial time variation can be hidden by assuming a restrictive prior structure. The posteriors for the other hyperparameters are basically unchanged.\(^\text{22}\)

\(^{22}\)In appendix F, we show empirical results for three additional VAR specifications: one that extends
CHOOSING PRIOR HYPERPARAMETERS

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.

Primcieri’s original dataset, a monthly VAR for the US and a VAR for international stock returns.
APPENDIX A: THE METROPOLIS-WITHIN-GIBBS STEP IMPLEMENTATION IN MORE DETAIL

In this section, we present details on the algorithm for the posterior sampling of the scaling parameters in the VAR with time-varying parameters and stochastic volatility. For the sake of brevity, we describe the sampling procedure for a generic scaling factor $k_X$, $X \in \{\Omega_b, \Omega_a, \Omega_h\}$. Given a draw for $X$, the conditional posterior $p(k_X|X) \propto p(X|k_X)p(k_X)$ can be obtained with a Metropolis-Hastings step. We use a version of the (Gaussian) random walk Metropolis-Hastings algorithm with an automatic tuning step for the proposal variance in a burn-in phase. The algorithm is initialized with values $k^0_X$ (which we choose to be the values from Primiceri (2005)) and $\sigma^2_{k_X}$, which we change in a preliminary burn-in phase to achieve a target acceptance rate.

1. At step $i$, take a candidate draw $k^*_X$ from $N(k^{i-1}_X, \sigma^2_{k_X})$

2. Calculate the acceptance probability $\alpha^i_{k_X} = \min\left(1, \frac{p(X|k^*_X)p(k^*_X)}{p(X|k^{i-1}_X)p(k^{i-1}_X)}\right)$

3. Accept the candidate draw by setting $k^i_X = k^*_X$ with probability $\alpha^i_{k_X}$. Otherwise set $k^i_X = k^{i-1}_X$.

4. Calculate the average acceptance ratio $\bar{\alpha}_{k_X}$. Adjust the increment standard deviation $\sigma_{k_X}$ every $q$th iteration according to $\sigma^{New}_{k_X} = \sigma_{k_X} \frac{\bar{\alpha}_{k_X}}{\alpha^*}$, where $\alpha^*$ denotes the target average acceptance ratio. Do not adjust after the iteration $i$ exceeds the burn-in threshold $I$.

In practice, we set $\alpha^* = .5$ and the burn-in threshold $I$ equal to one-half of the total repetition number.
APPENDIX B: THE ALGORITHM FOR A FIXED-COEFFICIENT VAR

Fixed coefficient VARs are often estimated using the Gibbs sampler (see Koop & Korobilis (2010)). A fixed coefficient Gaussian VAR is of the form:

\[(B.1)\quad y_t = \mu + \sum_{j=1}^{L} B_j y_{t-j} + e_t\]

with \(e_t \sim iid N(0, \Sigma)\).

If we define \(\beta \equiv [\mu' \; vec(B_1)' \ldots vec(B_L)']'\), the most commonly used Gibbs sampler assumes that

\[(B.2)\quad \beta \sim N(\beta(\phi), V_\beta(\phi))\]
\[(B.3)\quad \Sigma \sim IW(\nu V, \nu)\]

where we have made the dependence of the prior for \(\beta\) on hyperparameters \(\phi\) explicit. Note that the priors on \(\beta\) and \(\Sigma\) are assumed independent and are thus not natural conjugate priors (i.e. the approach of Giannone et al. (2013) cannot be applied in this case). We could also introduce additional hyperparameters for the prior on \(\Sigma\), but since popular priors such as the Minnesota prior focus on \(\beta\), we will do the same here. A Gibbs sampler for this model consists of the following three steps:

1. Draw \(\beta|\Sigma, \phi\)
2. Draw \(\Sigma|\beta, \phi\) - since this step conditions on \(\beta\), it simplifies to drawing \(\Sigma\) conditional only on \(\beta\) since \(\phi\) does not carry any additional information about \(\Sigma\) once we condition on \(\beta\)
3. Draw \(\phi|\beta, \Sigma\). As discussed in this paper, this simplifies to drawing \(\phi|\beta\)

The first two steps of the Gibbs sampler are standard in the literature (see again Koop & Korobilis (2010)), except that we have to possibly change \(\phi\) at every iteration when drawing \(\beta\). The last step is described in detail in this paper.
APPENDIX C: THE COMPLETE ALGORITHM FOR A TIME-VARYING PARAMETER VAR WITH STOCHASTIC VOLATILITY

In this section, we describe the complete algorithm to estimate the TVP-VAR model with stochastic volatility described in the main text. We modify the algorithm described in Del Negro & Primiceri (2015) to include additional steps for the drawing of the hyperparameters. The algorithm proceeds as follows:\footnote{A superscript $T$ denotes a sample of the relevant variable from $t = 1$ to $T$.}

1. Draw $h^T$ from $p(h^T|y^T, b^T, a^T, V, s^T, k_{\Omega_b}, k_{\Omega_a}, k_{\Omega_h})$. This step requires us to generate draws from a nonlinear state space system. We use the approach by Kim et al. (1998) to approximate draws from the desired distribution. For a correct posterior sampling of the stochastic volatilities, we follow the corrigendum in Del Negro & Primiceri (2015) and the modified steps therein.

2. Draw $b^T$ from $p(b^T|y^T, a^T, h^T, V, k_{\Omega_b}, k_{\Omega_a}, k_{\Omega_h})$. Conditional on all other parameter blocks, equations (4) and (5) from the main text form a linear Gaussian state space system. This step can be carried out using the simulation smoother detailed in Carter & Kohn (1994).

3. Draw $a^T$ from $p(a^T|y^T, b^T, h^T, V, k_{\Omega_b}, k_{\Omega_a}, k_{\Omega_h})$. Again, we draw these covariance states based on the simulation smoother of the previous step, exploiting our assumption that the covariance matrix of the innovations in the law of motion for the $a$ coefficients is block diagonal. This assumption follows Primiceri (2005), where further details on this step can be found.

4. Draw $\Omega_{h_b}$, $\Omega_{h_a}$, and $\Omega_{h_h}$. Given our distributional assumptions, these conditional posteriors of the time-invariant variances follow inverse-Wishart distributions (which are functions of $k_{\Omega_b}, k_{\Omega_a}, k_{\Omega_h}$).

5. Draw $s^T$, the sequence of indicators for the mixture of normals needed for the Kim et al. (1998) stochastic volatility algorithm.

6. Draw $k_{\Omega_b}, k_{\Omega_a}, k_{\Omega_h}$. Each of these scaling parameters is drawn via the algorithm described in section A of the appendix.
Next, we give a schematic overview of both the standard algorithm due to Del Negro & Primiceri (2015) and our extension.

**Algorithm 1** Standard TVP-VAR estimation procedure

1. Constant VAR on $T_0$ to initialize priors,
   set $g \sim 1$, where $g = 1, \ldots, G$
2. Sample $p(h^T | a^T, b^T, s^T, \Omega_b, \Omega_a, \Omega_h, y^T) = p_{FFBS}$
3. Sample $p(b^T | a^T, h^T, \Omega_b, \Omega_a, \Omega_h, y^T) = p_{FFBS}$
4. Sample $p(a^T | b^T, h^T, \Omega_b, \Omega_a, \Omega_h, y^T) = p_{FFBS}$
5. Sample $p(s^T | a^T, h^T, \Omega_b, \Omega_a, \Omega_h, y^T) \propto q \cdot f_N$
6. Sample $p(\Omega_b, \Omega_h, \Omega_a | \ldots) = p_{\Omega b}$
7. set $g \sim g + 1$ and go to **Step 2**.
   Iterate through **Step 2** and **Step 6** for large $G$ until convergence is achieved.

**Algorithm 2** Benchmark TVP-VAR with hyperparameters

1. Constant VAR on $T_0$ to initialize priors,
   set $g \sim 1$, where $g = 1, \ldots, G$
2. Sample $p(h^T | a^T, b^T, s^T, \Omega_b, \Omega_a, \Omega_h, y^T) = p_{FFBS}$
3. Sample $p(b^T | a^T, h^T, \Omega_b, \Omega_a, \Omega_h, y^T) = p_{FFBS}$
4. Sample $p(a^T | b^T, h^T, \Omega_b, \Omega_a, \Omega_h, y^T) = p_{FFBS}$
5. Sample $p(s^T | a^T, h^T, \Omega_b, \Omega_a, \Omega_h, y^T) \propto q \cdot f_N$
6. Sample $p(\kappa_{\Omega_b} | \Omega_b) = p(\Omega_b | \kappa_{\Omega_b})p(\kappa_{\Omega_b})$

   Sample $p(\kappa_{\Omega_h} | \Omega_h) = p(\Omega_h | \kappa_{\Omega_h})p(\kappa_{\Omega_h})$

   Sample $\prod_{j=1}^J p(\kappa_{\Omega_a} | \Omega_{a,j}) = p(\Omega_a | \kappa_{\Omega_a})p(\kappa_{\Omega_a})$
7. set $g \sim g + 1$ and go to **Step 2**.
   Iterate through **Step 2** and **Step 6** for large $G$ until convergence is achieved.
APPENDIX D: DENSITIES FOR HYPERPARAMETERS

In this section, we provide formulas for the prior densities that we use for the hyperparameters as well as additional results for alternative Monte Carlo settings.

D.1. Inverse Gamma Distribution

Our inverse gamma parameterization corresponds to the so called scaled inverse chi-squared distribution with scale parameter \( \tau \) and degree of freedom \( \nu \). Suppose \( x \) is inverse gamma distributed given shape \( \alpha \) and mode \( m \). Its scale parameter \( \beta \) is then \( m(\alpha + 1) \). The corresponding parameterization in the scaled inverse chi-squared specification is

\[
\begin{align*}
\nu &= 2\alpha \\
\tau &= \frac{2\beta}{\nu}
\end{align*}
\]

The density of the scaled inverse chi-squared distribution is given by

\[
f(x) = \frac{(\tau \nu/2)^{\nu/2} \exp \left(-\frac{\nu \tau^2}{2x}\right)}{\Gamma(\nu/2) x^{1+\nu/2}}
\]

D.2. Half-t And Half-Cauchy Distribution

The density of the half-t distribution is

\[
f(x) = 2\Gamma\left(\frac{\nu+1}{2}\right)\sqrt{\nu\pi\tau^2} \left[ 1 + \frac{1}{\nu} \left(\frac{x}{\tau}\right)^2 \right]^{-(\nu+1)/2}
\]

where \( \tau \) is the scale parameter and \( \nu \) the degree of freedom. For \( \nu = 1 \) the half-t distribution corresponds to the half-cauchy distribution and has the density function:

\[
f(x) = \frac{2\tau}{\pi(\tau^2 + x^2)}
\]

Both half-t and half-cauchy distribution do not have finite moments. Gelman (2006) proposes the half-t distribution as weakly informative prior for the standard deviation instead of the more conventional inverse gamma distribution.
Next, we carry out Monte Carlo simulations where we, instead of using our benchmark values for the fixed hyperparameters case, use a value of \( k_{\Omega} \) that is ten times higher. Tables E.1 to E.2 show that even with a substantially higher value of the fixed hyperparameter, our approach still does better.\textsuperscript{24}

### TABLE E.1
Monte Carlo results for \( DGP_4 \) with deterministic parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>( \mu_t )</th>
<th>( \phi_t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>i( \mathcal{G} )</td>
<td>0.7029</td>
<td>0.6780</td>
</tr>
<tr>
<td>half-Cauchy</td>
<td>0.6487</td>
<td>0.6495</td>
</tr>
<tr>
<td>half-t</td>
<td>0.7372</td>
<td>0.7405</td>
</tr>
<tr>
<td>Fixed</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

To show the effects of larger fixed hyperparameters, figure E.1 shows a version of figure ??, with the fixed hyperparameter \( k_{\Omega} \) increased by a factor of 10 (again, using a new set of 100 simulations, so the inverse-Gamma based results are not numerically identical to those in the main text, but very similar).

\textsuperscript{24}These Monte Carlo results were obtained from samples independent from the samples used in they main text, so the ratio of any two entries other than entries in the last column will not be numerically the same as they are in the corresponding tables in the main text, but they are numerically very close.
TABLE E.2
Monte Carlo results for $DGP_5$ with random walk parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\mu_t$</th>
<th>$\phi_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>i$\mathcal{G}$</td>
<td>0.8559</td>
<td>0.8694</td>
</tr>
<tr>
<td>half-Cauchy</td>
<td>0.8471</td>
<td>0.8818</td>
</tr>
<tr>
<td>half-t</td>
<td>0.8855</td>
<td>0.8961</td>
</tr>
<tr>
<td>Fixed</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Relative RMSE
[In-sample fit of parameter paths $\theta_t$ evaluated at posterior median]

<table>
<thead>
<tr>
<th>Horizons</th>
<th>i$\mathcal{G}$</th>
<th>half-Cauchy</th>
<th>half-t</th>
<th>Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9898</td>
<td>0.9614</td>
<td>0.9932</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.9732</td>
<td>0.9498</td>
<td>0.9657</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.9953</td>
<td>0.9931</td>
<td>0.9993</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.9829</td>
<td>0.9730</td>
<td>0.9797</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>0.9840</td>
<td>0.9754</td>
<td>0.9880</td>
<td>1.000</td>
</tr>
<tr>
<td>6</td>
<td>0.9783</td>
<td>0.9674</td>
<td>0.9697</td>
<td>1.000</td>
</tr>
<tr>
<td>7</td>
<td>0.9934</td>
<td>0.9907</td>
<td>0.9977</td>
<td>1.000</td>
</tr>
<tr>
<td>8</td>
<td>1.0026</td>
<td>1.0080</td>
<td>1.0029</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Relative RMSFE
[Out-of-sample forecast of $y_t$ evaluated at posterior median]

<table>
<thead>
<tr>
<th>Horizons</th>
<th>i$\mathcal{G}$</th>
<th>half-Cauchy</th>
<th>half-t</th>
<th>Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9898</td>
<td>0.9614</td>
<td>0.9932</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.9732</td>
<td>0.9498</td>
<td>0.9657</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.9953</td>
<td>0.9931</td>
<td>0.9993</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.9829</td>
<td>0.9730</td>
<td>0.9797</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>0.9840</td>
<td>0.9754</td>
<td>0.9880</td>
<td>1.000</td>
</tr>
<tr>
<td>6</td>
<td>0.9783</td>
<td>0.9674</td>
<td>0.9697</td>
<td>1.000</td>
</tr>
<tr>
<td>7</td>
<td>0.9934</td>
<td>0.9907</td>
<td>0.9977</td>
<td>1.000</td>
</tr>
<tr>
<td>8</td>
<td>1.0026</td>
<td>1.0080</td>
<td>1.0029</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Figure E.1.**—Estimated coefficient paths for the deterministic law of motion for parameters and the various priors for the hyperparameters in comparison.
APPENDIX F: ADDITIONAL EMPIRICAL RESULTS

As additional evidence, we provide the posterior distribution of hyperparameters to three additional VAR models fitted to data sets that differ in the frequency, time period, dimension and type of data. For the Primiceri VAR we quarterly data on year-over-year inflation, the unemployment rate and the Federal Funds rate from the second quarter of 1948 to the first quarter of 2014 to update Primiceri’s dataset. For our monthly VAR, we use US data on the growth rate of industrial production, the Federal Funds rate, year-over-year inflation rate and stock return data from March 1972 to January 2014. The stock-return VAR uses monthly data on real returns of stock market indices for Canada, Italy, the US, the UK and Germany from March 1972 to January 2014. We use two lags for each VAR. In figure F.1 we compare those to the fixed hyperparameter in Primiceri (2005). Each row reports the results for a given data set and each column refers to one of the four hyperparameters (we focus here on the case in which intercepts are governed by a separate hyperparameter). Depending on the specific data set the posterior distribution can be more dispersed or peaked. Its posterior mean and can be significantly to the left indicating evidence of a lower extent of time variation and vice versa. For the case of $\kappa_{\Omega_h}$ in all data sets the posterior distribution has a mean one order of magnitude higher than indicated by the fixed value employed in Primiceri (2005). Conversely, the posterior distribution of $\kappa_{\Omega_a}$ is in all empirical applications peaks substantively to the left of the value chosen by Primiceri. Hence, the a-priori extent of time variation of the covariance states is strongly rejected by the data indicating less time variation. For the hyperparameters related to the coefficient states $\kappa_{\Omega_b,intercept}$ and $\kappa_{\Omega_b,coefficient}$ the evidence is mixed. Not surprisingly, the posterior distribution is reasonably close to the fixed choices of Primiceri for the data set that is similar to the one used in his paper. Note again that he employed a model choice exercise based on RJMCMC to pick the hyperparameters. There is nonetheless evidence across data sets for different values. This evidence further reiterates that our data-driven approach is a useful extension to the estimation of time-varying parameter models.
Figure F.1.— Posterior distribution of each hyperparameter for three additional VARs employed.
REFERENCES


Korobilis, D. (2014), Data-based priors for vector autoregressions with drifting coefficients, MPRA Paper 53772, University Library of Munich, Germany.


