Mateus Rigo Noriller

Maximum likelihood estimation of a TVP-VAR

Florianópolis, SC

2018

Mateus Rigo Noriller

MAXIMUM LIKELIHOOD ESTIMATION OF A TVP-VAR

Dissertação de mestrado em conformidade com as normas ABNT apresentado à comissão avaliadora como requisito parcial para a obtenção de título de mestre em economia.
Orientador: Guilherme Valle Moura

Florianópolis, SC
2018

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Noriller, Mateus Rigo
Maximum likelihood estimation of a TVP-VAR / Mateus Rigo Noriller ; orientador, Guilherme Valle Moura, 2018.
64 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico, Programa de Pós-Graduação em Economia, Florianópolis, 2018.

Inclui referências.

1. Economia. 2. Vetores autorregressivos. 3. Coeficientes variantes. 4. Volatilidade estocástica multivariada. 5. Máxima verossimilhança. I. Moura, Guilherme Valle. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Economia. III. Título.

Mateus Rigo Noriller

Maximum likelihood estimation of a TVP-VAR

Trabalho aprovado. Florianópolis, SC, 19 de março de 2018:

Prof. Dr. Jaylson Jair da Silveira
Coordenador de curso - UFSC/PPGEco

Prof. Dr. Guilherme Valle Moura
Orientador - UFSC/PPGEco

Banca examinadora:

Prof. Dr. André Alves Portela Santos
Membro interno 1 - UFSC/PPGEco

Prof. Dr. Francis Carlo Petterini Lourenço
Membro interno 2 - UFSC/PPGEco

Prof. Dr. Hedibert Freitas Lopes
Membro externo - INSPER (videoconferência)

Florianópolis, SC
2018

ABSTRACT

This master's thesis proposes a vector autoregression with time-varying coefficients and multivariate stochastic volatility which can be estimated in one step by maximum likelihood. The coefficients are assumed to follow random walks and the volatility of the system is modelled as a Wishart process, increasing the flexibility in describing the behavior of stochastic covariances. Exploiting the conjugacy between Normal, Wishart and multivariate beta distributions, filtering formulas for tracking the latent states are derived in closed form. The method is then applied to U.S. data and generates results that are similar to those reported in the macroeconomic literature.

Keywords: Vector autoregressions. Time-varying coefficients. Wishart stochastic volatility. Maximum likelihood.

RESUMO

Esta dissertação de mestrado propõe um modelo vetor autorregressivo com coeficientes variantes e volatilidade estocástica multivariada que pode ser estimado por máxima verossimilhança. Supõe-se que os coeficientes sigam passeios aleatórios e modela-se a volatilidade do sistema através de um processo Wishart, que confere flexbilidade na descrição de covariâncias estocásticas. Ao explorar a conjugação entre as distribuições normal, Wishart e beta multivariada, são derivadas fórmulas de filtragem em forma fechada que permitem rastrear os estados não-observáveis. O modelo é, então, utilizado para analisar dados dos Estados Unidos e gera resultados similares aos reportados na literatura macroeconômica.

Palavras-chave: Vetores autorregressivos. Coeficientes variantes. Volatilidade estocástica tipo Wishart. Máxima verossimilhança.

CONTENTS

INTRODUCTION

In empirical macroeconomics, the coefficients of vector autoregressions can be thought of as summarizing structural economic relationships. In the context of the Phillips curve, for example, the coefficients of the equation for interest rate can be interpreted as describing a monetary policy rule (COGLEY; SARGENT, 2005). In this sense, traditional vector autoregressions with constant coefficients rest on the assumption that economic relationships are stable over time (KOOP; KOROBILIS, 2010; DEL NEGRO; SCHORFHEIDE, 2011). They are thus suited to describing economic behavior that is approximately linear and does not exhibit substantial variation (LUBIK; MATTHES, 2015).

However, given that most macroeconomic time series do exhibit some form of nonlinearity, the assumption of constant coefficients might be too restrictive in many applications. For example, it is a well-known stylized fact about the U.S. economy that inflation was more volatile and persistent in the 1970s (the Great Inflation) than in the 1980s and subsequently (the Great Moderation) (KOOP; KOROBILIS, 2010; LUBIK; MATTHES, 2015). The fact that, at different times, macroeconomic variables seem to behave differently suggests that the structure of the true stochastic process underlying the economy is of a dynamical nature. In other words, economic relationships seem to change over time and this is consistent with the fundamental notion that economies experience fluctuations to which agents are not oblivious.

To capture such structural changes in economic relationships, econometricians developed different types of vector autoregressions with time-varying coefficients. Broadly speaking, such models can be divided into two classes: one in which coefficients vary gradually over time according to an autoregressive process and another in which they change abruptly as in Markov-switching or structural-break models (DEL NEGRO; SCHORFHEIDE, 2011). This master's thesis focuses on the former class and we shall refer to models belonging to it simply as *time-varying parameter VARs* (TVP-VARs).

Cogley & Sargent (2001) were responsible for popularizing VARs with autoregressive coefficients in macroeconomics. In the debate on U.S. monetary policy, these authors maintained that the Great Inflation of the 1970s was the result of bad policy. Their argument ran along the following lines. The way the FED reacted to inflation changed over time because of its changing views about the existence of an exploitable trade-off between inflation and unemployment. In other words, under different chairmanships the FED would demonstrate different levels of willingness to inflate the economy. As the monetary authority executes its (changing) monetary policy, agents learn an imperfect version of the Phillips curve and gradually adjust their expectations and decisions. This gradual adjustment constitutes a structural change in economic relationships. Consequently, the "bad policy" version of the story relies on evidence that the VAR coefficients changed over time (KOOP; KOROBILIS, 2010). This motivated Cogley & Sargent (2001) to propose a homoscedastic VAR with drifting coefficients.

However, their model was immediately criticized for ignoring another potential source of nonlinearity: time-variation in the variances and covariances of the VAR innovations. In particular, Sims (2001) and Stock (2001) argued that a more appropriate specification would be a vector autoregression with constant coefficients and stochastic volatility. This criticism is closely related to the "bad luck" version of the story, according to which the worse inflation-unemployment outcomes of the 1970s resulted from more volatile exogenous shocks. More importantly, it led to the development of models that could account for and distinguish both phenomena, i.e., models with autoregressive coefficients *and* multivariate stochastic volatility (MSV).

The distinguishing features of these TVP-VARs are the structure of multivariate stochastic volatility and the properties of the shocks to the coefficients. The two most popular approaches are the ones proposed by Cogley & Sargent (2005) and Primiceri (2005). They assume homoscedastic shocks to the coefficients and specify the MSV as a set of univariate stochastic volatilities based on the model of Jacquier et al. (1994). These models can capture rich dynamics of multiple time series in a flexible way, but statistical inference in a nonlinear state-space setting usually requires simulation techniques and can be quite challenging. Cogley & Sargent (2005) use the single-move Gibbs sampling algorithm, which is relatively simple. Nonetheless, their specification is restrictive because the covariances of the shocks to the economy are not allowed to change independently. The model of Primiceri (2005) is more flexible and does not impose this restriction. For this reason, however, it requires a more complicated multi-move Gibbs sampler.

The goal of this master's thesis is to propose a TVP-VAR that does not impose restrictions on the evolution of innovation covariances and that, at the same time, can be estimated in a more straightforward way. To do so, we build upon the work of Uhlig (1997) and propose a TVP-VAR with Wishart stochastic volatility that has exact filtering formulas. Then we borrow insights from Kim (2014) and derive an analytical expression for the likelihood function. As a result, the coefficients and the stochastic volatility can be tracked by means of a recursive algorithm and the unknown parameters can be estimated in one step by maximum likelihood.

The text is organized as follows. Chapter 1 establishes a theoretical background on state-space models and TVP-VARs. It offers just enough content to enable the reader to proceed to chapter 2, where Wishart TVP-VARs are discussed. In section 2.1 we present the original model of Uhlig (1997) and analyze some of its properties. Section 2.2 discusses our extended version of Uhlig's model with drifting coefficients. Then section 2.3 establishes the results that constitute the most important contribution of this work: exact filtering formulas and a likelihood function. Finally, in chapter 3 we apply our method to a dataset for the U.S. economy which is similar to the one used in Cogley & Sargent (2005). As will be seen, our model captures time-variation in the VAR coefficients and in the volatility of the shocks to the economy. In general, the results resemble those already presented in the macroeconomic literature.

1 BACKGROUND

1.1 STATE-SPACE MODELS AND THE FILTERING PROBLEM

Many economic phenomena can be described as dynamical systems and can thus be analyzed using state-space models. In this approach, it is assumed that the development over time of the system under study is determined by unobservable vectors of states. They are associated with observable measurements which convey information about the states with some degree of imperfection and the relation between them is specified by the state-space model. The central purpose of state-space analysis is to infer the important properties of the states from a knowledge of the observations (DURBIN; KOOPMAN, 2012). There are different ways of doing that. This section introduces the basic ideas of one of them: filtering.

It is often the case that measurements of certain variables become available periodically. This implies that the states of a system would have to be reestimated every time a new observation was received. However, storing data and reprocessing known information can be computationally costly. The so-called recursive filters offer a solution to this problem by processing new information and by reestimating the states sequentially. They consist of basically two steps: prediction and update. In the former, the researcher utilizes a model of the system and all available information to make the best possible prediction about its states. When a new piece of information becomes available, the latter step uses Bayes' theorem to modify the initial prediction and to update the knowledge of the researcher about the states of the system (RISTIC et al., 2004).

Let t be a discrete time index. The vector of states, denoted by α_t, is not observable and evolves according to the following stochastic model:

$$\alpha_t = f_{t-1}(\alpha_{t-1}, \upsilon_{t-1}), \tag{1.1}$$

where f_{t-1} is a known function and υ_{t-1} represents a sequence of perturbations (noise) to the vector of states. Let y_t be a vector of observable measurements. The vectors y_t and α_t are related through the following measurement equation:

$$y_t = h_t(\alpha_t, \varepsilon_t), \tag{1.2}$$

where h_t is a known function and ε_t is a sequence of errors in the measurements. ε_t and υ_{t-1} are assumed to be white noises, with known probability density functions and mutually independent. In addition to that, it is assumed that the PDF of the initial states vector, $p(\alpha_1)$, is known and is independent of the errors for all periods of time.

In filtering problems, the goal is to estimate α_t recursively using the collected measurements y_t. Let Y_t denote the set of all measurements up to period t, i.e., $Y_t = \{y_i, i = 1, ..., t\}$. Likewise, let Y_{t-1} denote the set of measurements available up to period $t-1$. From a Bayesian viewpoint, the problem is to quantify the degree of confidence that one has about the value taken by state α_t given the information available up to period t. In other words, the researcher seeks to build the posterior distribution $p(\alpha_t|Y_t)$ and this can be done recursively through the prediction and the update steps (RISTIC et al., 2004).

Assuming that the PDF $p(\alpha_{t-1}|Y_{t-1})$ is known, the prediction step uses model (1.1) to obtain the prior distribution (or the prediction density) for the states vector in period t through the Chapman-Kolmogorov equation:

$$p(\alpha_t|Y_{t-1}) = \int p(\alpha_t|\alpha_{t-1})p(\alpha_{t-1}|Y_{t-1})d\alpha_{t-1}. \tag{1.3}$$

Note that the term $p(\alpha_t|\alpha_{t-1})$ is the probabilistic model of the evolution of the states vector and is determined by (1.1). In period t, when a new measurement y_t becomes available, the update step *updates* the prediction PDF (1.3) via Bayes' rule:

$$\begin{aligned} p(\alpha_t|Y_t) &= p(\alpha_t|y_t, Y_{t-1}) \\ &= \frac{p(y_t|\alpha_t, Y_{t-1})p(\alpha_t|Y_{t-1})}{p(y_t|Y_{t-1})} \\ &= \frac{p(y_t|\alpha_t)p(\alpha_t|Y_{t-1})}{p(y_t|Y_{t-1})}. \end{aligned} \tag{1.4}$$

The expressions above use the fact that $p(\alpha_t|\alpha_{t-1},Y_{t-1}) = p(\alpha_t|\alpha_{t-1})$ and $p(y_t|\alpha_t,Y_{t-1}) = p(y_t|\alpha_t)$ because, by definition, α_t contains all relevant information about the system. The term $p(y_t|\alpha_t)$ is determined by the measurement model (1.2). It also appears in the normalizing constant in the denominator of (1.4):

$$p(y_t|Y_{t-1}) = \int p(y_t|\alpha_t)p(\alpha_t|Y_{t-1})d\alpha_t. \tag{1.5}$$

To sum up, the update step uses newly available measurements to modify the predictive distribution and obtain a posterior density which summarizes the updated knowledge about the current states. These two steps solve the problem of exact and complete characterization of the updated distribution in a recursive manner. For this reason, they are said to form the basis of the optimal solution to the filtering problem. Then the updated distribution enables the researcher to obtain estimates related to the vector of states. However, this is a conceptual solution which, in general, cannot be determined analytically. Only in a few particular cases it can be characterized exactly and completely by a finite, fixed and sufficient statistic. In such cases, optimal algorithms are employed to deduce this solution (RISTIC et al., 2004). For example, in the context of a linear and Gaussian structure the prediction and update steps produce the Kalman filter. The model of Uhlig (1997) (and ours, for that matter) is another example: by choosing transition equations and prior distributions appropriately, it is possible to use the prediction and the update steps to derive an exact, recursive algorithm that characterizes completely the posterior distribution of the states in each period.

Here, and probably elsewhere in this text, we employ a Bayesian terminology for the sake of convenience. It should be noted, however, that filtering and state-space models are by no means confined to the realm of Bayesian statistics. The method described above is sometimes referred to as a Bayesian filter because of the Bayes' rule employed in the update step, but it is just as widely used in classical statistics [see, for example, Harvey (1993) and Durbin & Koopman (2012)].

As we shall see, vector autoregressions with time-varying parameters fit very conveniently into a state-space setting. The measurement model (1.2) takes the form of a VAR equation, while the dynamic parameters are treated as unobservable states with given laws of motion which play the role of the transition equation (1.1).

In practice, state-space models contain not only unobservable states, but also unknown parameters that need to be estimated. For example, the variances of the error terms above, σ_ε^2 and σ_υ^2, are hardly ever known. Durbin & Koopman (2012) show that state estimates are the same whether classical or Bayesian analysis is employed. Parameters, on the other hand, require different treatments. In classical analysis, they are assumed to be fixed but unknown, whereas in Bayesian analysis observations are assumed to be fixed and parameters are interpreted as random variables (DURBIN; KOOPMAN, 2012).

Estimating the unknown parameters by maximum likelihood is a conceptually straightforward matter. When initial conditions are known, as is the case above, the likelihood of the entire sample can be factored as:

$$L(Y_T) = p(y_1, ..., y_T) = p(y_1) \prod_{t=2}^{T} p(y_t | Y_{t-1}), \tag{1.6}$$

where T denotes the last time period for which measurements are available. Each term on the rightmost side of (1.6) is a normalizing constant, given by (1.5), which depends on the parameters of the model. Once the researcher has figured out the form of $p(y_t|Y_{t-1})$, it is easy to assemble (1.6) and to maximize it with respect to the parameters. This is the approach that we will follow in this paper.

On the other hand, Bayesian estimation of the parameters is based on posterior analysis and, in virtually all cases of interest, requires simulation methods. This paper will not cover the theory behind that for the sake of brevity. The interested reader is referred to chapter 13 of Durbin & Koopman (2012) or to West & Harrison (1997) for a general Bayesian treatment of state-space models. In particular, Koop & Korobilis (2010) and Del Negro & Schorfheide (2011) provide good introductions to Bayesian TVP-VARs.

1.2 TVP-VAR MODELS IN MACROECONOMICS

Vector autoregressions with time-varying coefficients were popularized in macroeconomics by Cogley & Sargent (2001). Their goal was to analyze the changing behavior of the inflation-unemployment-interest-rate dynamics in the United States after World War II. To do so, they proposed a trivariate VAR with coefficients that are random walks. Written in state-space form, their model can be represented by the following equations:

$$y_t = B_t X_t + \mathcal{U}(H^{-1})'\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_k), \quad (1.7)$$

$$\operatorname{vec}(B'_{t+1}) = \operatorname{vec}(B'_t) + \mathcal{U}(\mathcal{Q}^{-1})'\eta_t, \qquad \eta_t \sim \mathcal{N}(0, I_{kl}), \quad (1.8)$$

where $t = 1, ..., T$ denotes time and y_t is a k-dimensional vector of endogenous variables observed at time t. $X_t := [C'_t \ y'_{t-1} \ \cdots \ y'_{t-p}]'$, where C_t is a c-dimensional vector of deterministic regressors, such as intercept and trend. $B_t := [B_{0,t} \ B_{1,t} \ \cdots \ B_{p,t}]$. $B_{0,t}$, which is $(k \times c)$, and $B_{j,t}$ for $j = 1, ..., p$, which is $(k \times k)$, are both coefficients matrices. For $l := c + kp$, X_t is $(l \times 1)$ and B_t is $(k \times l)$. H is a $(k \times k)$ symmetric, positive definite precision matrix and $\mathcal{U}(.)$ stands for its upper Cholesky factor. All ε_t, $t = 1, ..., T$, are k-dimensional vectors of independently distributed exogenous shocks and $\mathcal{N}(0, I_k)$ denotes the multivariate normal distribution. As for the transition equation, $\operatorname{vec}(.)$ is the operator that stacks the columns of a matrix, $\mathcal{Q}$ is a $(kl \times kl)$ positive definite, symmetric precision matrix and the η_t are also kl-dimensional vectors of independently distributed exogenous shocks.

As far as the parameters are concerned, the model of Cogley & Sargent (2001) contains only one source of time-variation: the matrix of coefficients, B_t. The rationale for drifting coefficients is that agents learn about the economy as it changes and news arrive. As a result, they adapt their decision rules to the new circumstances and this causes the coefficients in B_t to change in unpredictable ways (COGLEY; SARGENT, 2001). Clearly, this interpretation is intimately related to the LUCAS's (1976) critique. And to the extent that agents have an imperfect view of things, this learning process occurs gradually and can thus be captured by models in which the coefficients change from period to period (as opposed to regime-switch models) (DEL NEGRO; SCHORFHEIDE, 2011). In the U.S. inflation-unemployment debate, "new circumstances" can be read as changes in the transmission mechanism of monetary policy due to changing beliefs of the monetary authority about the existence of exploitable trade-offs between inflation and unemployment (KOOP; KOROBILIS, 2010). Hence, finding empirical evidence that the VAR coefficients drifted over time would translate into evidence that inflation behavior changed because of changes in the monetary authority's view about the Phillips curve. This is the "bad policy" story sustained by Cogley & Sargent (2001) and others.

Cogley & Sargent (2001) did find evidence of parameter drift for the U.S. economy. However, their model was criticized by Sims (2001) and Stock (2001) who questioned the assumption of a fixed, non-stochastic covariance matrix H^{-1}. These authors pointed to the evidence presented by Bernanke & Mihov (1998a), Bernanke & Mihov (1998b) and others that the innovation variances had changed over time, whereas the VAR coefficients had remained stable. As a matter of fact, if the world were characterized by constant B and drifting H^{-1}, fitting a model with constant H^{-1} and drifting B could cause the estimates of B to drift simply to compensate for the misspecification of H^{-1}, thus exaggerating the time-variation in B (COGLEY; SARGENT, 2005). According to the "bad luck" version of the story, the distribution of the exogenous shocks evolved, but agents' responses to them did not.

To tackle these criticisms, Cogley & Sargent (2005) extended their earlier model by allowing both B and H to vary and adopted a multivariate version of the stochastic volatility model of Jacquier et al. (1994):

$$H_t^{-1} = A^{-1} R_t A^{-1'}, \tag{1.9}$$

where

$$R_t = \begin{bmatrix} r_{1t} & 0 & 0 \\ 0 & r_{2t} & 0 \\ 0 & 0 & r_{3t} \end{bmatrix}. \tag{1.10}$$

The diagonal elements of R_t are independent, univariate stochastic volatilities that evolve as driftless, geometric random walks:

$$\ln r_{it} = \ln r_{it-1} + \sigma_i \psi_{it}, \qquad \psi_{it} \sim \mathcal{N}(0,1), \tag{1.11}$$

where the ψ_{it} represent volatility innovations and the σ_i are scaling parameters that determine their magnitudes. Another key assumption in this model is:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ \alpha_{21} & 1 & 0 \\ \alpha_{31} & \alpha_{32} & 1 \end{bmatrix}. \tag{1.12}$$

This matrix of fixed parameters is the one which, loosely speaking, determines the covariances between innovations. Its unique elements, the α_{ij}'s, do not have time subscripts because they are assumed to be constant over time. From (1.10) and (1.12), one can easily read off the product in (1.9) and conclude that the covariances between the shocks to the economy, the off-diagonal elements in H_t^{-1}, are allowed to vary over time, but only in a tightly restricted fashion: as fixed proportions of the innovations' variances (KOOP; KOROBILIS, 2010). This constitutes a major drawback in the model of (COGLEY; SARGENT, 2005) since it is too restrictive in important applications. In impulse-response analysis, for instance, a constant A matrix implies that an innovation to the i-th variable has a time-invariant effect on the j-th variable (PRIMICERI, 2005).

At about the same time, Primiceri (2005) proposed a similar but more flexible model which extends (1.12) to the time-varying case:

$$A_t = \begin{bmatrix} 1 & 0 & 0 \\ \alpha_{21_t} & 1 & 0 \\ \alpha_{31_t} & \alpha_{32_t} & 1 \end{bmatrix}, \tag{1.13}$$

where the unrestricted elements, $\alpha_t := [\alpha_{21_t}\ \alpha_{31_t}\ \alpha_{32_t}]'$, evolve according to

$$\alpha_t = \alpha_{t-1} + \zeta_t, \qquad \zeta_t \sim \mathcal{N}(0, D). \tag{1.14}$$

In principle, (1.13) implies that the the evolution of innovations' covariances is not restricted in any way (KOOP; KOROBILIS, 2010). In his empirical analysis, however, Primiceri (2005) assumes that D is block diagonal, with blocks corresponding to parameters of separate equations. This means that the shocks to the covariances (between the innovations to the economy) are independent across equations, or equivalently, that the coefficients of contemporaneous relations among variables evolve independently in each VAR equation. (PRIMICERI, 2005) shows that this assumption is not crucial in his model. Relaxing it, however, does make estimation and inference more complicated (KOOP; KOROBILIS, 2010).

Later on, motivated by the results of Stock & Watson (2007), Cogley et al. (2010) extended their previous models by allowing stochastic volatility in the parameter innovations as well, i.e., by letting $\mathcal{Q}_t$ in (1.8) change over time. The law of motion of $\mathcal{Q}_t$ in Cogley et al. (2010) mimics that of H_t in Cogley & Sargent (2005) described by (1.9)-(1.12), which simplifies estimation, but implies that the evolution of covariances is tied to that of variances.

A homoscedastic TVP-VAR such as that of Cogley & Sargent (2001) constitutes a normal, linear state-space model. As such, Bayesian inference about objects of interest is easy to deal with by means of relatively simple MCMC methods (KOOP; KOROBILIS, 2010). In particular, various efficient algorithms have been developed to allow for posterior simulation of B_t, for $t = 1, ..., T$, conditional on the other unknown parameters. Carter & Kohn (1994) and Frühwirth-Schnatter (1994) are two prominent examples.

When multivariate stochastic volatility is introduced, statistical inference becomes more involved. Univariate stochastic volatility models treat dynamic variances as stochastic processes and are amenable to forward filtering and backward sampling. This means that one can easily take joint draws of the latent states conditional upon the data and the parameters (WINDLE; CARVALHO, 2014). Replicating this property in the multivariate case, though, can be challenging. In particular, it is difficult to construct a reasonable matrix-valued stochastic process that respects positive definiteness and at the same time couples nicely to the observation distribution. Positive definiteness can be ensured by means of transformations which define the process in different coordinate systems. However, these transformations tend to make state-space inference more complicated because the product of the observation and the transition densities does not usually yield a recognizable posterior distribution which can be easily simulated (WINDLE; CARVALHO, 2014).

In the case of the TVP-VARs explained above, the introduction of multivariate stochastic volatility implies that additional algorithms within the MCMC routine are needed. Cogley & Sargent (2005) construct a Metropolis-within-Gibbs sampler and, in particular, they use the univariate algorithm of Jacquier et al. (1994) to sample stochastic volatilities. This suffices for their specification, which restricts the evolution of innovation covariances. The model of Primiceri (2005), on the other hand, is more complex and thus requires a different approach. He simulates the latent covariance matrices from their posterior distributions with the Gibbs sampling method of Kim et al. (1998), which consists in transforming a nonlinear and non-Gaussian state space form into a linear and approximately Gaussian one (PRIMICERI, 2005).

The bottom line is that Bayesian estimation of TVP-VARs becomes more complicated as the flexibility of the model increases. Loosely speaking, each time one allows a given parameter to be time-varying, an extra layer of complexity is added to the estimation method. So much so that the algorithm of Primiceri (2005) contained an error which was later corrected by Del Negro & Primiceri (2015). Although the results remained qualitatively similar, the fact that the model was used in many applications and the flaw in the Gibbs implementation remained undetected for almost a decade can be seen as an indication of the complexity of the estimation procedure. These difficulties can be overcome by adopting the MSV structure proposed by Uhlig (1997). The next chapter explores it in detail, but the gist of Uhlig's approach is a clever choice for the volatility transition equation, one which allows forward filtering (and backward sampling) to be done in closed form and makes statistical inference easier even within a Bayesian framework.

2 WISHART TVP-VAR

2.1 THE MODEL OF UHLIG (1997)

Uhlig (1997) proposes a multivariate version of the local scale model of Shephard (1994):

$$y_t = BX_t + \mathcal{U}(H_t^{-1})'\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_k), \tag{2.1}$$

$$H_{t+1} = \frac{1}{\lambda}\mathcal{U}(H_t)'\Theta_t\mathcal{U}(H_t), \qquad \Theta_t \sim \mathcal{B}_k\left(\frac{v+l}{2}, \frac{1}{2}\right), \tag{2.2}$$

where B is a matrix of constant coefficients and Θ_t, $t = 1, ..., T$, are $(k \times k)$, symmetric, positive definite and independently distributed matrices of innovations to the precision. Equation (2.1) is identical to (1.7), except that it assumes fixed, non-stochastic coefficients. $\lambda > 0$ and $v > k - 1$ are parameters and $\mathcal{B}_k$ represents the multivariate beta distribution.

As $v \to \infty$, the multivariate beta distribution for Θ_t concentrates around the identity matrix and the model approaches a homoscedastic TVP-VAR. As such, the parameter v governs the degree of time-variation in the precision through the multiplicative beta innovations: the smaller v is, the more the process H_t fluctuates (UHLIG, 1997). λ also controls the speed at which the precision moves, but it does so directly. Given v, the process H_t can be asymptotically degenerate or explosive depending on whether λ is too large or too small, respectively (KIM, 2014). In practice, λ also controls how the model attributes weights to different observations when forming estimates and one-step ahead predictions for the measurements (WINDLE; CARVALHO, 2014).

This model has a few desirable properties. First of all, the Wishart distribution guarantees that H_t is always positive definite. Secondly, Θ_t is a complete matrix in which all elements vary stochastically. This means that the variances and covariances of the shocks to the economy change over time without restrictions, and there is no need to model them separately. In particular, the evolution of covariances is not tied to the evolution of variances, as is the case in Cogley & Sargent (2005), and no assumptions are made about the covariances being hit by shocks that are independent across variables, as is the case in the model implemented by Primiceri (2005). Finally, due to the conjugacy between the Wishart and the multivariate singular beta distributions established in Uhlig (1994), the model possesses closed form filtering formulas.

The uninitiated but careful reader might be wondering why (2.2) is referred to as a Wishart process. Consider two independent, Wishart-distributed matrices: $A \sim \mathcal{W}_k(n_1, \Sigma)$ and $B \sim \mathcal{W}_k(n_2, \Sigma)$, with $n_1 > k-1$ and $n_2 > k-1$. Define the following Choleski decomposition: $A+B = H = \mathcal{U}(H)'\mathcal{U}(H)$. Let Θ be a $k \times k$ symmetric matrix such that $A = \mathcal{U}(H)'\Theta\mathcal{U}(H)$. Then Theorem 3.3.1 and Definition 3.3.2 of Muirhead (1982) establish that $H \sim \mathcal{W}_k(n_1+n_2, \Sigma)$ and $\Theta \sim \mathcal{B}_k(n_1/2, n_2/2)$. This defines de multivariate beta distribution. Once this is done, the same results can be read backwards: assuming $\Theta \sim \mathcal{B}_k(n_1/2, n_2/2)$ and $H \sim \mathcal{W}_k(n_1+n_2, \Sigma)$, as we will in this paper, it follows that $A = \mathcal{U}(H)'\Theta\mathcal{U}(H) \sim \mathcal{W}_k(n_1, \Sigma)$. Hence (2.2) defines a Wishart process.

As a matter of fact, when Uhlig mentions the "conjugacy between the Wishart and the multivariate beta distributions", he does not mean it in the usual Bayesian sense of combining a prior with a likelihood and obtaining a posterior which belongs to the same family of distributions as the prior. Instead, what he has in mind is the result presented above: interpret the Wishart-distributed matrix H as a prior, allow it to be hit by multiplicative beta innovations and what you get at the other end is a Wishart posterior.

It is important to notice, however, that in (2.2) we have $n_2 = 1$, which does not satisfy the restriction $n_2 > k-1$. Theorem 1 of Uhlig (1994) establishes exactly this backward-reading fact for the case in which n_2 is allowed to be any positive integer. Equivalently, this theorem extends the previous definition of the multivariate beta distribution to the general case. Then in Theorem 7 Uhlig (1994) derives the probability density function of this newly defined random matrix for the singular case, when $n_2 = 1$. And because the Wishart distribution is fundamental to the study of the multivariate beta distribution, Uhlig (1994) also extends the former to the general case of $n_2 > 0$. Later on, Diaz-Garcia & Jaimez (1997) derived the PDF of the multivariate beta distribution for the general case.

Then the question becomes: why does Uhlig (1997) specify $n_2 = 1$? Let us do a non-rigorous imagination exercise with a simplified version of the model (the skeptical reader will be satisfied in section 2.3 and appendix C). Assume that the researcher starts with the following prior: $H_t|Y_{t-1} \sim \mathcal{W}_k(v, v^{-1}S_{t-1}^{-1})$, so that $\mathrm{E}(P_{t-1})^{-1} = S_{t-1}$. Assume also that he or she observes a single measurement vector such that $y_t \sim \mathcal{N}(0, H_t^{-1})$. Then the researcher can update his or her knowledge about H_t through Bayes' rule, as explained in section 1.1. By doing so, the resulting updated knowledge will be that $H_t|Y_t \sim \mathcal{W}_k(v+1, (v+1)^{-1}S_t^{-1})$, for S_t of a certain form. Thus the updating process added one degree of freedom to the prior and altered its scale matrix. As a matter of fact, this will happen in every update step of this problem.

Now it is time to evolve H_t one step forward. $H_t|Y_t$ is interpreted as the new prior and the goal is to find the predictive density $H_{t+1}|Y_t$. Here comes the crucial fact: we, and Uhlig, are interested in an algorithm which preserves recursiveness and conjugacy, so that we can obtain exact solutions to the filtering problem. As such, in $t+1$ we want a predictive density which is of the same form as the one in period t. So we know that $H_{t+1}|Y_t$ will have to be Wishart-distributed. But we also know that, eventually, we will be able to update our knowledge about H_{t+1} and that the update step will add one degree of freedom to this Wishart distribution. Therefore, the law of motion for the precision has to yield a Wishart posterior and has to eat away one degree of freedom from the prior, which is Wishart with $v+1$ d.f. (otherwise we would run into trouble in the next update step). Precisely, the transition equation for the precision has to yield $H_{t+1}|Y_t \sim \mathcal{W}_k(v, v^{-1}S_t^{-1})$. By inspecting Theorem 1 of Uhlig (1994), or by reading backwards Theorem 3.3.1 of Muirhead (1982), we come to the conclusion that this can only happen if we specify that $\Theta_{t+1} \sim \mathcal{B}_k(v/2, 1/2)$, i.e., if we set $n_2 = 1$.

In the model of Uhlig (1997), the precision matrix evolves with the same frequency with which observations are received: once in every time step. For this reason, observations are said to have rank 1 and the trick of setting $n_2 = 1$ suffices to solve the problem described above. However, there are cases in which this synchrony between observations and precision does not hold. For example, a common case in financial econometrics is the one in which the researcher is interested in estimating the variance of daily returns, assuming it changes only once a day, but observes multiple intraday vectors of returns. I.e., the precision matrix changes less often than the frequency with which measurements are observed. In that case, the observation matrix would have rank $r > 1$ and it would be necessary to set $n_2 = r$. Windle & Carvalho (2014) explore this set-up and extend the model of Uhlig (1997) to the general case of rank $r \geq 1$.

To sum up, the choice of the distribution for Θ_t is the one that facilitates the filtering problem. If we assume that $H_t|Y_t$ has an acceptable distribution, then we need a transition equation that yields $H_{t+1}|Y_t$ with an appropriate distribution to update, so that $H_t|Y_t$ will have a distribution that lets us repeat the process (WINDLE; CARVALHO, 2014).

The model of Uhlig (1997) is constructed as a Bayesian vector autoregression, which means that inference is based on posterior analysis and requires the elicitation of a prior. Uhlig chooses a prior distribution of a specific form so as to allow for a flexible treatment of unit roots. But this choice has a consequence: the resulting posterior is proportional to a Normal-Wishart distribution scaled with a function which depends on the coefficients in B. Hence, even though the model produces exact filtering formulas, inference about the states requires the posterior distribution to be evaluated numerically. Uhlig (1997) then employs importance-sampling.

For philosophical or practical reasons, though, one might be willing to depart from Bayesian estimation. It turns out that Uhlig's model is flexible enough to accommodate that: Kim (2014) showed that the analytical filtering formulas can be used to derive an exact likelihood function so that the parameters can be estimated in one step by maximum likelihood. We shall follow this approach in the next sections.

2.2 AN EXTENSION OF UHLIG'S MODEL

The previous section showed that Uhlig's model has a few desirable properties: it guarantees positive definiteness for the precision without imposing restrictions on covariances, it possesses closed-form filtering formulas and it is amenable to classical estimation. Nonetheless, it relies on the critical assumption that the VAR coefficients are time-invariant. In this section, we extend Uhlig's MSV-VAR to allow for drifting coefficients in the tradition of Cogley & Sargent (2001). Because we emphasize maximum likelihood estimation, our model can also be seen as an extension of Kim (2014). Ultimately, it is built in such a way that the desirable properties of the original model carry over to a full-fledged TVP-VAR.

We assume conditionally Gaussian observations, as in equation (1.7), but model the precision according to equation (2.2), as described above:

$$\begin{aligned} y_t &= B_t X_t + \mathcal{U}(H_t^{-1})'\varepsilon_t, & \varepsilon_t &\sim \mathcal{N}(0, I_k), \\ H_{t+1} &= \frac{1}{\lambda}\mathcal{U}(H_t)'\Theta_t\mathcal{U}(H_t), & \Theta_t &\sim \mathcal{B}_k\left(\frac{v+l}{2}, \frac{1}{2}\right). \end{aligned} \tag{2.3}$$

To extend the constant-coefficient VAR of Uhlig (1997) to the time-varying case, we assume that the coefficients follow an autoregressive process similar to the one given by equation (1.8), but decompose $\mathcal{Q}_t$ into two terms connected by the Kronecker product:

$$\text{vec}(B'_{t+1}) = \text{vec}(B'_t) + [\mathcal{U}(H_t^{-1}) \otimes \mathcal{U}(Q^{-1})]'\eta_t, \quad \eta_t \sim \mathcal{N}(0, I_{kl}), \tag{2.4}$$

where Q is $(l \times l)$, symmetric and positive definite and η_t, $t = 1, ..., T$, are kl-dimensional vectors of independently distributed innovations. By defining γ_t such that

$$\text{vec}(\gamma'_t) = [\mathcal{U}(H_t^{-1}) \otimes \mathcal{U}(Q^{-1})]'\eta_t, \tag{2.5}$$

the transition equation (2.4) can be equivalently represented in matrix form as:

$$B_{t+1} = B_t + \gamma_t, \qquad \gamma_t | H_t \sim \mathcal{N}_{k,l}(0_{k,l}, H_t^{-1} \otimes Q^{-1}), \tag{2.6}$$

where $\mathcal{N}_{k,l}$ denotes the $(k \times l)$-dimensional matrix variate normal distribution and $0_{k,l}$ is a $(k \times l)$ matrix of zeros [see Definition 2.2.1 of Gupta & Nagar (2000)]. We shall stick to this representation because it relates more easily to the model of Uhlig (1997). Such a structure implies that:

$$B_{t+1}|B_t, H_t \sim \mathcal{N}_{k,l}(B_t, H_t^{-1} \otimes Q^{-1}). \tag{2.7}$$

As can be seen in (2.6)-(2.7), the covariance matrix of the innovations, H_t^{-1}, is also a source of variation in the evolution of the VAR coefficients, i.e., the shocks to the coefficients are themselves heteroscedastic.

The Kronecker product structure means that the covariances between the shocks in γ_t have a column-specific component, H_t, as well as a row-specific component, Q. Then H_t^{-1} represents the covariances between the shocks to each VAR equation (or variable). Similarly, Q^{-1} denotes the covariances between the l shocks to any row of B_t (or to the coefficients of any VAR equation). Take any two elements of B_t, e.g. the intercepts of the first and second VAR equations: $B_{0,t}^1$ and $B_{0,t}^2$. Each of them is hit by a different shock, say $\gamma_{0,t}^1$ and $\gamma_{0,t}^2$. Then the Kronecker product implies that $cov(\gamma_{0,t}^1, \gamma_{0,t}^2) = h_t^{1,2} q^{0,0}$, where $h_t^{1,2}$ represents the covariance between the shocks that hit variables 1 and 2, and $q^{0,0}$ denotes the covariance between the shocks to the intercept (or in this case the variance) of any given row of B_t. One can think of Q^{-1} and H_t^{-1} as the coefficient-specific and the variable-specific components of the covariance matrix of γ_t, respectively.

Such an assumption has important implications. As is usual in a TVP-VAR, the observations in y_t can increase or decrease because of changes in B_t and because the shocks to the economy suddenly become more or less volatile through H_t. Here, because the covariance matrix of γ_t is tied to H_t, the coefficients in B_t themselves will also become more or less volatile. From an economic perspective, this is equivalent to assuming that more abrupt changes in the economy due to more volatile exogenous shocks will cause agents to re-optimize their decision rules more aggressively. In this respect, we depart from Cogley & Sargent (2005) and Primiceri (2005), whose models assume homoscedastic innovations to the VAR coefficients, and propose a more general specification in line with Cogley et al. (2010) and with the empirical findings of Stock & Watson (2007). As a matter of fact, if the justification for drifting coefficients presented in Cogley & Sargent (2001) is coherent enough, then the assumption of heteroscedastic shocks to the coefficients is theoretically more plausible than its homoscedastic counterpart.

2.3 EXACT FILTERING FORMULAS

This section presents the most important contribution of this thesis. The following propositions and corollary establish the results needed to track the unobservable states and to estimate the unknown parameters. Their proofs are presented in Appendix C. Notation-wise, let $Y_t := \{y_{1-p}, \cdots, y_t\}$ denote the set of all available measurements up to time t. Assume that y_t obeys equation (2.3) for all t, that H_t and B_t evolve according to (2.2) and (2.6), respectively, and that at period t all past observations are known. Additionally, let $\mathcal{NW}_{k,l}$ and $\mathcal{T}_{k,l}$ denote the matrix normal-Wishart and matrix variate t distributions, respectively. Appendix B contains definitions of these distributions. Implicitly condition on the parameters v, λ and Q and assume the following initial condition: $B_1, H_1|Y_0 \sim \mathcal{NW}_{k,l}(B_{1|0}, N_{1|0}, v, S_{1|0})$, with $B_{1|0}$, $N_{1|0}$ and $S_{1|0}$ known and $N_{1|0}$ and $S_{1|0}$ symmetric.

Proposition 1.
Suppose $B_t, H_t|Y_{t-1} \sim \mathcal{NW}_{k,l}(B_{t|t-1}, N_{t|t-1}, v, S_{t|t-1})$, *where* $N_{t|t-1}$ *and* $S_{t|t-1}$ *are positive definite and symmetric. After observing* y_t, *the filtered joint density of the states is given by*

$$B_t, H_t|Y_t \sim \mathcal{NW}_{k,l}(B_{t|t}, N_{t|t}, v+1, S_{t|t}), \tag{2.8}$$

where:

$$B_{t|t} = (B_{t|t-1}N_{t|t-1} + y_t X_t')N_{t|t}^{-1}, \tag{2.9}$$

$$N_{t|t} = N_{t|t-1} + X_t X_t', \tag{2.10}$$

$$S_{t|t} = \frac{v}{v+1} S_{t|t-1} + \frac{1}{v+1} e_t (1 - X_t' N_{t|t}^{-1} X_t) e_t', \tag{2.11}$$

$$e_t = y_t - B_{t|t-1} X_t. \tag{2.12}$$

Additionally, the filtered marginal densities are given by

$$B_t | Y_t \sim \mathcal{T}_{k,l}[v - k + 2, B_{t|t}, (v+1) S_{t|t}, N_{t|t}^{-1}], \tag{2.13}$$

$$H_t | Y_t \sim \mathcal{W}_k[v+1, (v+1)^{-1} S_{t|t}^{-1}]. \tag{2.14}$$

Equations (2.9) and (2.11) are the ones we are most interested in. They represent the conditional expected values of the marginal distributions (see Appendix B). As such, (2.9) can be directly interpreted as filtered estimates for the coefficients and (2.11) as the filtered estimates for the volatility of the system.

The next thing we want to do is obtain the one-step-ahead predictive density. We use the normal-Wishart PDF from the previous theorem, together with the normal-Wishart density implied by the transition equations, to integrate out the states of period t analytically. The resulting predictive density is also normal-Wishart.

Proposition 2.
Suppose $B_t, H_t | Y_t \sim \mathcal{NW}_{k,l}(B_{t|t}, N_{t|t}, v+1, S_{t|t})$*, where* $N_{t|t}$ *and* $S_{t|t}$ *are positive definite and symmetric. Then the predictive joint density of the states is given by*

$$B_{t+1}, H_{t+1} | Y_t \sim \mathcal{NW}_{k,l}(B_{t+1|t}, N_{t+1|t}, v, S_{t+1|t}), \tag{2.15}$$

where:

$$B_{t+1|t} = B_{t|t}, \tag{2.16}$$

$$N_{t+1|t} = [Q^{-1} + (\lambda N_{t|t})^{-1}]^{-1}, \tag{2.17}$$

$$S_{t+1|t} = \lambda \frac{v+1}{v} S_{t|t}. \tag{2.18}$$

Additionally, the predictive marginal densities are

$$B_{t+1} | Y_t \sim \mathcal{T}_{k,l}(v - k + 1, B_{t+1|t}, v S_{t+1|t}, N_{t+1|t}^{-1}), \tag{2.19}$$

$$H_{t+1} | Y_t \sim \mathcal{W}_k(v, v^{-1} S_{t+1|t}^{-1}). \tag{2.20}$$

Note that (2.15) has the same form as the initial condition required by Theorem 1. So, paraphrasing Uhlig (1994), the filtering game can begin anew.

The intuition underlying the proofs of Theorems 1 and 2 is actually quite simple: we combine two normal-Wishart densities analytically and as a result we get another normal-Wishart density. This process is analogous to what is done with Gaussian distributions in the Kalman filter.

So both the updated and the predictive densities are known and have known properties. But in all these results we are implicitly conditioning on the unknown parameters which still need to be estimated. The next result establishes what is needed to estimate them by maximum likelihood. It is actually a corollary of Theorem 1 because the likelihood contribution of period t is the integrating constant of the updated density, which in this case is a normal-Wishart. It is stated here for the sake of reference and because it is relevant on its own for estimation:

Corollary 1.
Suppose $B_t, H_t|Y_{t-1} \sim \mathcal{NW}_{k,l}(B_{t|t-1}, N_{t|t-1}, v, S_{t|t-1})$, *where* $N_{t|t-1}$ *and* $S_{t|t-1}$ *are positive definite and symmetric. Then the period-t likelihood contribution is*

$$y_t|Y_{t-1} \sim t_k[v-k+1, B_{t|t-1}X_t, (v-k+1)^{-1}\Sigma_t^{-1}], \tag{2.21}$$

where:

$$\Sigma_t = (1 - X_t' N_{t|t}^{-1} X_t)(vS_{t|t-1})^{-1}, \tag{2.22}$$

$$N_{t|t} = N_{t|t-1} + X_t X_t'. \tag{2.23}$$

Note that t_k denotes the usual multivariate t distribution. The likelihood of the entire sample can be factored as: $f(Y_T) = \prod_{t=1}^{T} f(y_t|Y_{t-1})$ (DURBIN; KOOPMAN, 2012). Then from Theorem 1 it follows that the log-likelihood function is given by

$$\log[f(Y_T)] = -\frac{Tk}{2}\log[(v+1-k)\pi] + T\log\left[\Gamma\left(\frac{v+1}{2}\right)\right]$$

$$-T\log\left[\Gamma\left(\frac{v+1-k}{2}\right)\right]+\frac{1}{2}\sum_{t=1}^{T}\log(|(v+1-k)\Sigma_t|)$$
$$-\frac{(v+1)}{2}\sum_{t=1}^{T}\log[1+(y_t-B_{t|t-1}X_t)'\Sigma_t(y_t-B_{t|t-1}X_t)], \qquad (2.24)$$

where $|.|$ represents the determinant of a matrix and $\Gamma(.)$ denotes the gamma function.

To sum up, we have derived an algorithm with exact filtering formulas that allow us to track the unobservable states as well as an analytical expression for the likelihood function. This means that we can estimate the parameters of the model by maximum likelihood in one step without resorting to simulation techniques.

3 EMPIRICAL ANALYSIS

In order to visualize what our method is capable of, we apply it to a dataset for the U.S. economy which mimics that of Cogley & Sargent (2005) and contains the following time series:

- Short-term nominal interest rate: secondary market rate on 3-month Treasury bills. The data are sampled monthly and then converted to a quarterly series by selecting the first month of each quarter.

- Unemployment rate: civilian unemployment rate. The original series is seasonally adjusted and sampled monthly. It is converted to a quarterly series by taking the average of monthly rates. Within the VAR, we use the logit transformation of the unemployment rate to ensure that its expectations lie between zero and one (COGLEY et al., 2010).

- Inflation: Consumer Price Index for all urban consumers, all items. The original series is seasonally adjusted and sampled monthly. It is converted to a quarterly series by point-sampling the third month of each quarter. Inflation is then measured as the log-differences of these values.

The data are available from the Federal Reserve Economic Database (FRED) and have codes TB3MS, UNRATE and CPIAUCSL, respectively. The relevant sample spans the period 1948:Q2 to 2000:Q4 and comprises 211 quarterly observations for each variable.

Both Uhlig (1997) and Kim (2014) work with pre-set values for v and λ. As noted by Kim (2014), these parameters bear no economic meaning and, in the case of v, it is debatable whether estimating it yields better results relative to when it is pre-set. Uhlig (1997) suggests using $v = 20$ for quarterly data and $\lambda = v/(v+1)$ to allow for a reasonable degree of time-variation in the precision. Notice from (2.18) that $\lambda = v/(v+1)$ implies $E[H_{t+1}^{-1}|Y_t] = S_{t|t}$, which is analogous to the random walk behavior of B_t. However, based on results reported in the literature, we have an a priori idea of what the volatility of the system should look like in the analyzed period. After conducting sensitivity tests with respect to these parameters, we concluded that $v = 20$ and $\lambda = v/(v+1)$ do not introduce enough variability to capture the expected movements in the volatility. It turns out that much better results can be achieved by working with lower values. As such, we pre-set $v = 10$ and $\lambda = 0.8$. In addition to that, we work with two lags, which is a common choice in the TVP-VAR literature, and include only an intercept in each equation. Thus for $c = 1$, $p = 2$ and $k = 3$, the parameters that need to be estimated correspond to the 28 unique elements of the precision matrix Q in (2.6).

To start recursions, we must pick initial values for some elements of the model. To construct $B_{1|0}$, we follow Uhlig (1997) and use the estimates of a constant-coefficient VAR with one lag. $N_{1|0}$ represents, loosely speaking, our degree of confidence in this choice for $B_{1|0}$. We want this to be a weak choice whose effect dies out quickly, so we build $N_{1|0}$ as a matrix with diagonal elements equal to 0.001 and zeros elsewhere. For $S_{1|0}$, we also follow Uhlig (1997) and estimate three separate AR(1) models, one for each variable, and save the squared residuals. $S_{1|0}$ is then built as a diagonal matrix whose diagonal entries are the averages of these squared residuals.

And to start optimization, we also need initial values for Q. Recall that Q^{-1} is interpreted as the covariance matrix of the shocks to the coefficients of any VAR equation (i.e., the coefficient-specific component of the volatility). We fit several VAR(1) models with an expanding window (ideally, if sample-size permits, one would use a rolling window). Then we use the estimated matrices of coefficients, $\hat{B}_t$, to compute $\hat{\gamma}_t$ from (2.6) for various periods. This gives us 3 rows of residuals (one for the coefficients of each VAR equation) with 7 elements in each (because we have 7 coefficients per equation):

$$\hat{\gamma}_t = \begin{bmatrix} \hat{\gamma}_{1,t} \\ \hat{\gamma}_{2,t} \\ \hat{\gamma}_{3,t} \end{bmatrix} = \begin{bmatrix} \hat{\gamma}^0_{1,t} & \hat{\gamma}^{1,1}_{1,t} & \hat{\gamma}^{1,2}_{1,t} & \hat{\gamma}^{1,3}_{1,t} & \hat{\gamma}^{2,1}_{1,t} & \hat{\gamma}^{2,2}_{1,t} & \hat{\gamma}^{2,3}_{1,t} \\ \hat{\gamma}^0_{2,t} & \hat{\gamma}^{1,1}_{2,t} & \hat{\gamma}^{1,2}_{2,t} & \hat{\gamma}^{1,3}_{2,t} & \hat{\gamma}^{2,1}_{2,t} & \hat{\gamma}^{2,2}_{2,t} & \hat{\gamma}^{2,3}_{2,t} \\ \hat{\gamma}^0_{3,t} & \hat{\gamma}^{1,1}_{3,t} & \hat{\gamma}^{1,2}_{3,t} & \hat{\gamma}^{1,3}_{3,t} & \hat{\gamma}^{2,1}_{3,t} & \hat{\gamma}^{2,2}_{3,t} & \hat{\gamma}^{2,3}_{3,t} \end{bmatrix}.$$

The superscripts on the rightmost hand-side of the equality above are to be read as follows: $\hat{\gamma}_{i,t}^{0}$ refers to the innovation to the intercept of variable i and $\hat{\gamma}_{i,t}^{l,j}$ denotes the estimated shock to the coefficient of the l-th lag of variable j in the equation of variable i. For each row, we calculate the variances of the estimated residuals across the various subsamples. This gives us a measure of how the shocks to the coefficients in each row change as we add more observations. Then for each of the 7 coefficients, we take the averages of these variances across the 3 rows. This, in turn, gives us a measure of how the innovations to the coefficients of *any given row* change. Finally, we build Q_0 with the *inverses* of these averages in the diagonal entries (because they denote precision, not variance) and with zeros elsewhere.

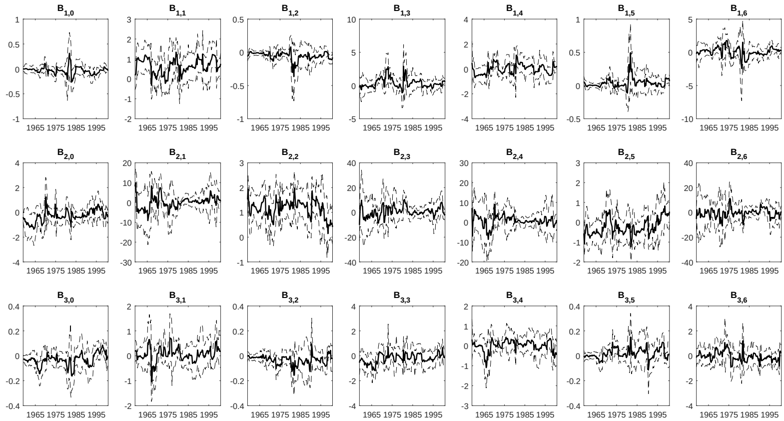
Figures 1 and 2 present the evolution of the filtered coefficients, $E[B_t|Y_t] = B_{t|t}$. Figure 2 is directly comparable to the lower portion of figure 2 of Cogley & Sargent (2005). The dashed lines in figure 1 are 95% confidence intervals and each row in it represents one VAR equation, with

$$y_t = \begin{bmatrix} interest_t \\ unemployment_t \\ inflation_t \end{bmatrix}.$$

In constructing the confidence intervals, we used Theorem 4.3.1 of Gupta & Nagar (2000) to compute the marginal filtered variances of $B_t|Y_t$.

Some coefficients change more pronouncedly than others and several of them can be regarded as stable over time, especially considering the confidence intervals. Qualitatively, these results are similar to those of Cogley & Sargent (2005). However, our method seems to capture more time-variation in the coefficients than theirs, probably due to the heteroscedasticity specified in the law of motion of these states in our model. In figure 2, we plot all filtered estimates together and it turns out that it is quite difficult to follow one line and make sense of what is happening. In contrast, the corresponding figure of Cogley & Sargent (2005) is much cleaner, with only two or three coefficients changing over time. This result reinforces the idea mentioned previously that the homoscedastic law of motion used in Cogley & Sargent (2005) and Primiceri (2005) might be too restrictive, as suggested by the results of

Figure 1 – Filtered coefficients (A).



Stock & Watson (2007).

Figure 2 – Filtered coefficients (B).

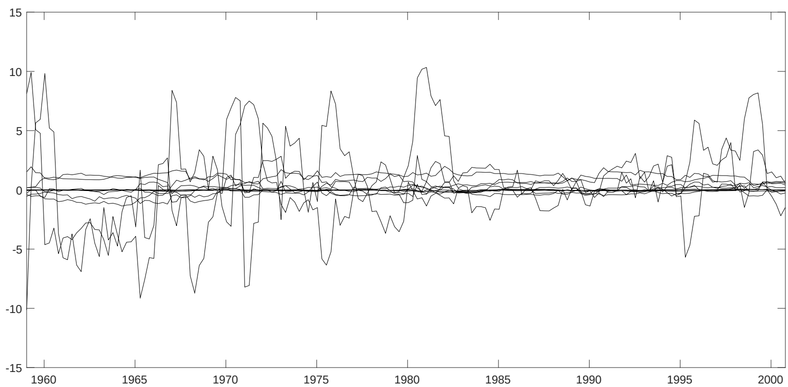
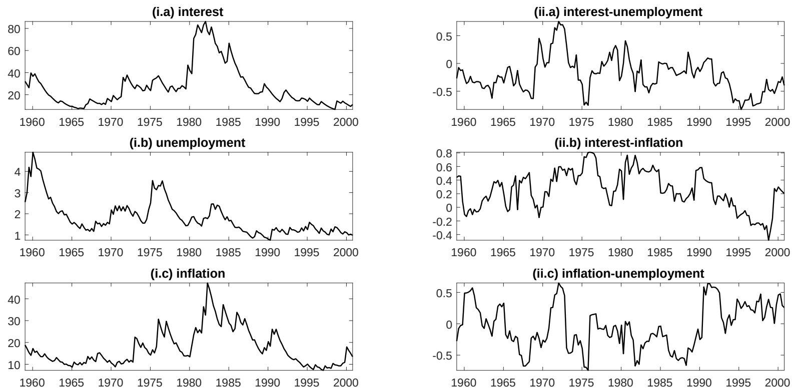


Figure 3 presents the evolution of the filtered volatility of the system, $E[H_t^{-1}|Y_t] = S_{t|t}$. It is directly comparable to figure 3 of Cogley & Sargent (2005). The first column portrays standard deviations of the shocks to each variable, with (i.a) and (i.c) being expressed in basis points. The second column shows correlation coefficients between pairs of innovations. Our model clearly captures time-variation in the volatility of the shocks. As far as the standard deviations are concerned, our results resemble those of Cogley & Sargent (2005). As for the correlation coefficients, the results are not so similar: while in Cogley & Sargent (2005) the correlations mimic the standard deviations, in our model the correlations evolve independently. In general, our model suggests a more smooth time-variation in the volatility of innovations.

Figure 3 – Filtered volatility.



4 CONCLUDING REMARKS

This master's thesis proposed a vector autoregression with drifting coefficients and multivariate stochastic volatility which can be estimated by maximum likelihood. It extended the model of Uhlig (1997) and borrowed insights from Kim (2014) for the new approach to estimation.

The VAR coefficients are modelled as an autoregressive process with heteroscedastic shocks. The covariance matrix of these innovations is decomposed into two terms connected by the Kronecker product. One of these terms, the coefficient-specific component, is fixed and non-stochastic. The other term, the variable-specific component, is stochastically time-varying and corresponds to the same covariance matrix of the shocks to the observations. In other words, coefficients and observations share a common source of volatility. An economic interpretation of this structure is offered along the lines of the original motivation for drifting coefficients.

The multivariate stochastic volatility is introduced in the form of a Wishart process in which the precision is hit by multiplicative beta innovations. The Wishart process respects positive definiteness and does not impose restrictions on the evolution of covariances. Most importantly, it couples analytically with the measurement equation and with the chosen law of motion for the coefficients. In particular, we benefit from the conjugacy results established in Uhlig (1994) and Uhlig (1997). They enabled us to derive closed-form filtering formulas for the latent states, as well as an analytical expression for the likelihood function. This means that the coefficients and the volatility of the system can be tracked by means of an exact algorithm and the unknown parameters can be estimated in one step by maximum likelihood. There is no need for simulation-based inference.

Therefore, the proposed method is flexible while retaining simplicity of estimation. We applied it to U.S. macroeconomic data and it was able to detect time-variation in the VAR coefficients as well as in the volatility of the innovations. In general, the results resemble those already presented in the macroeconomic literature.

Nonetheless, a word of caution is needed. Strictly speaking, our method in its current form does not constitute an alternative to the usual approach to TVP-VARs for it does not do exactly the same job. With Bayesian TVP-VARs, statistical inference is carried out via posterior analysis. Hence the knowledge of the researcher about B_j, for $j \in [1,T]$, is based on the entire data set [see, for example, equation 18 of Cogley & Sargent (2005)]. The same is true about other objects of interest. In contrast, the results presented above constitute a recursive *filter* aimed at period-by-period learning. As such, filtered estimates for B_j and H_j take into account the information available up to period j, but disregard all observations between periods $j+1$ and T (notice that estimates for the additional parameters do not suffer from this). This is a limitation in the sense that future observations can convey information about past and present states. For this reason, the next hurdle in our research agenda is the derivation of a smoothing algorithm also in closed form. We suspect this is possible by combining results from Windle & Carvalho (2014) with those of the Kalman filter.

BIBLIOGRAPHY

BAUWENS, L.; LUBRANO, M.; RICHARD, J. F. *Bayesian inference in dynamic econometric models*. [S.l.]: OUP Oxford, 1999. 51, 53, 59

BERNANKE, B. S.; MIHOV, I. The liquidity effect and long-run neutrality. In: ELSEVIER. *Carnegie-Rochester conference series on public policy*. [S.l.], 1998. v. 49, p. 149–194. 20

BERNANKE, B. S.; MIHOV, I. Measuring monetary policy. *The Quarterly Journal of Economics*, MIT Press, v. 113, n. 3, p. 869–902, 1998. 20

CARTER, C. K.; KOHN, R. On gibbs sampling for state space models. *Biometrika*, Oxford University Press, v. 81, n. 3, p. 541–553, 1994. 22

COGLEY, T.; PRIMICERI, G. E.; SARGENT, T. J. Inflation-gap persistence in the us. *American Economic Journal: Macroeconomics*, American Economic Association, v. 2, n. 1, p. 43–69, 2010. 22, 31, 35

COGLEY, T.; SARGENT, T. J. Evolving post world-war ii u.s. inflation dynamics. *NBER Macroeconomics Annual 2001*, v. 16, p. 331–388, 2001. 11, 12, 19, 20, 22, 29, 31

COGLEY, T.; SARGENT, T. J. Drifts and volatilities: monetary policies and outcomes in the post wwii us. *Review of Economic Dynamics*, v. 8, n. 2, p. 262–302, 2005. 11, 12, 13, 20, 21, 22, 23, 25, 30, 35, 37, 38, 39, 42

DEL NEGRO, M.; PRIMICERI, G. E. Time varying structural vector autoregressions and monetary policy: a corrigendum. *Review of Economic Studies*, v. 82, n. 4, p. 1342–1345, 2015. 23

DEL NEGRO, M.; SCHORFHEIDE, F. Bayesian macroeconometrics. In: ______. *The Oxford handbook of bayesian econometrics*. [S.l.]: Oxford Handbooks, 2011. p. 293–389. 11, 18, 20

DIAZ-GARCIA, J. A.; JAIMEZ, R. G. Proof of the conjectures of h. uhlig on the singular multivariate beta and the jacobian of a certain matrix transformation. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 25, n. 5, p. 2018–2023, 1997. 26

DURBIN, J.; KOOPMAN, S. J. *Time series analysis by state space methods*. 2. ed. Oxford: Oxford University Press, 2012. 15, 17, 18, 33

FRÜHWIRTH-SCHNATTER, S. Data augmentation and dynamic linear models. *Journal of time series analysis*, Wiley Online Library, v. 15, n. 2, p. 183–202, 1994. 22

GREENBERG, E. *Introduction to Bayesian econometrics*. 2. ed. [S.l.]: Cambridge University Press, 2008. 64

GUPTA, A.; NAGAR, D. Matrix variate distributions. In: BREZIS, H.; DOUGLAS, R. G.; JEFFREY, A. (Ed.). *Monographs and surveys in pure and applied mathematics*. London: Chapman & Hall/CRC, 2000. 30, 37, 51, 63

HARVEY, A. C. *Time series models*. 2. ed. [S.l.]: MIT Press, 1993. 17

JACQUIER, E.; POLSON, N. G.; ROSSI, P. E. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, v. 12, n. 4, 1994. 12, 20, 23

KIM, D. Maximum likelihood estimation for vector autoregressions with multivariate stochastic volatility. *Economics Letters*, v. 123, n. 3, p. 282–286, 2014. 13, 25, 28, 29, 35, 41

KIM, S.; SHEPHARD, N.; CHIB, S. Stochastic volatility: likelihood inference and comparison with arch models. *The review of economic studies*, Oxford University Press, v. 65, n. 3, p. 361–393, 1998. 23

KOOP, G.; KOROBILIS, D. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, Now Publishers, Inc., v. 3, n. 4, p. 267–358, 2010. 11, 12, 18, 20, 21, 22

LUBIK, T. A.; MATTHES, C. Time-varying parameter vector autoregressions: specification, estimation, and an application. *Economic Quarterly*, Federal Reserve Bank of Richmond, v. 101, n. 4, p. 323–352, 2015. 11

LUCAS, R. E. Econometric policy evaluation: a critique. In: BRUNNER, K.; MELTZER, A. (Ed.). *The Phillips curve and labor markets*. [S.l.]: North-Holland, 1976. v. 1, p. 19–46. 19

MUIRHEAD, R. J. *Aspects of multivariate statistical theory*. [S.l.]: John Wiley & Sons, Inc., 1982. 26, 27

PETERSEN, K. B.; PEDERSEN, M. S. The matrix cookbook. p. 72, 2012. 60

PRIMICERI, G. E. Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, v. 72, n. 3, p. 821–852, 2005. 12, 13, 21, 22, 23, 26, 30, 37

RISTIC, B.; ARULAMPALAM, S.; GORDON, N. *Beyond the Kalman filter*: particle filters for tracking applications. Boston: Artech House, 2004. 15, 16, 17, 55, 57

SHEPHARD, N. Local scale models: State space alternative to integrated garch processes. *Journal of Econometrics*, Elsevier, v. 60, n. 1-2, p. 181–202, 1994. 25

SIMS, C. A. Comment on cogley and sargent's 'evolving post-world war ii us inflation dynamics'. *NBER macroeconomics annual*, MIT Press, v. 16, p. 373–379, 2001. 12, 20

STOCK, J. H. Comment on cogley and sargent's 'evolving post-world war ii us inflation dynamics'. *NBER macroeconomics annual*, MIT Press, v. 16, p. 379–387, 2001. 12, 20

STOCK, J. H.; WATSON, M. W. Why has us inflation become harder to forecast? *Journal of Money, Credit and Banking*, v. 39, n. 1, p. 3–33, 2007. 22, 31, 38

UHLIG, H. On singular wishart and singular multivariate beta distributions. *The Annals of Statistics*, v. 22, n. 1, p. 395–405, 1994. 26, 27, 33, 41

UHLIG, H. Bayesian vector autoregressions with stochastic volatility. *Econometrica*, v. 65, n. 1, p. 59–73, 1997. 13, 17, 23, 25, 27, 28, 29, 30, 35, 36, 41, 51, 54, 58

WEST, M.; HARRISON, J. *Bayesian forecasting & dynamic models*. 2. ed. [S.l.]: Springer-Verlag, 1997. 18

WINDLE, J.; CARVALHO, C. M. A tractable state-space model for symmetric positive-definite matrices. *Bayesian Analysis*, International Society for Bayesian Analysis, v. 9, n. 4, p. 759–792, 2014. 22, 23, 25, 28, 42

Appendix

APPENDIX A – RELEVANT DIMENSIONS

Table 1 – Dimensions of vectors and matrices.

y_t	$k \times 1$
$B_{0,t}$	$k \times c$
$B_{j,t},\ j = 1,...,p$	$k \times k$
$B_t, B_{t\|t}, B_{t+1\|t}$	$k \times l$
X_t	$l \times 1$
C_t	$c \times 1$
ε_t	$k \times 1$
η_t	$kl \times 1$
γ_t	$k \times l$
Θ_t	$k \times k$
e_t	$k \times 1$
$H_t, S_{t\|t}, S_{t+1\|t}, \Sigma_t$	$k \times k$
$Q, N_{t\|t}, N_{t+1\|t}$	$l \times l$
$l := c + kp$	

APPENDIX B – SOME IMPORTANT DISTRIBUTIONS

The propositions presented in section 2.3 rely extensively on the definitions of two major distributions: the matrix normal-Wishart and the matrix variate t. The purpose of this appendix is to fix notation so as to avoid confusion. For a detailed account of matrix variate distributions, see (GUPTA; NAGAR, 2000). For a more succinct description, refer to Appendix A.2 of (BAUWENS et al., 1999).

Our definition of the matrix normal-Wishart distribution is based on Appendix A of (UHLIG, 1997). Assume the following matrices: B and $\bar{B}$ are $(k \times l)$, H is $(k \times k)$, symmetric and positive definite, S is $(k \times k)$ and positive definite, N is $(l \times l)$ and positive definite. Let $v \geq k > 0$. Then the probability density function of a matrix normal-Wishart distribution for B and H with parameters $\bar{B}$, N, v and S is given by:

$$f^{k,l}_{\mathcal{NW}}(B, H; \bar{B}, N, v, S) = \\ = \kappa_{\mathcal{NW}} \; |H|^{0.5(l+v-k-1)} exp\{-0.5\,\mathrm{tr}[((B-\bar{B})N(B-\bar{B})' + vS)H]\}, \tag{B.1}$$

where

$$\kappa^{k,l}_{\mathcal{NW}} = \frac{|N|^{0.5k} \, \left(\frac{v}{2}\right)^{0.5kv} |S|^{0.5v}}{(2\pi)^{0.5kl} \; \Gamma_k\left(\frac{v}{2}\right)}. \tag{B.2}$$

The normal-Wishart distribution results from the multiplication of the densities of a conditional matrix normal variable and a Wishart variable. Using some properties of the trace, it is easy to check that (B.1) implies:

$$B|H \sim \mathcal{N}_{k,l}(\bar{B}, H^{-1} \otimes N^{-1}), \tag{B.3}$$

$$H \sim \mathcal{W}_k[v, (vS)^{-1}]. \tag{B.4}$$

For details on the $\mathcal{N}_{k,l}$ and $\mathcal{W}_k$ distributions, see sections 2.2 and 3.2 of (GUPTA; NAGAR, 2000), respectively.

For the matrix variate t distribution, we adopt the definition of (GUPTA; NAGAR, 2000). Consider the same matrices described above and assume $v > 0$. The random matrix B $(k \times l)$ follows a matrix variate t distribution with parameters v, $\bar{B}$, S and N if its PDF is given by

$$f_{\mathcal{T}}^{k,l}(B;v,\bar{B},S,N) = \pi^{-0.5kl}\frac{\Gamma_k\left(\frac{v+l+k-1}{2}\right)}{\Gamma_k\left(\frac{v+k-1}{2}\right)}$$
$$|N|^{-0.5k}|S|^{-0.5l}|I_k + S^{-1}(B-\bar{B})N^{-1}(B-\bar{B})'|^{-0.5(v+l+k-1)}. \quad \text{(B.5)}$$

We denote this by $B \sim \mathcal{T}_{k,l}(v,\bar{B},S,N)$.

APPENDIX C – PROOFS OF PROPOSITIONS

First we establish an important result that will be used in the proofs of the propositions presented in section 2.3. It is a slight variation of Theorem A.19 of Bauwens et al. (1999).

Proposition 3.
Assume that B and H jointly follow a matrix normal-Wishart distribution with parameters $\bar{B}$, N, v and S, such that:

$$B|H \sim \mathcal{N}_{k,l}(\bar{B}, H^{-1} \otimes N^{-1}),$$
$$H \sim \mathcal{W}_k[v, (vS)^{-1}].$$

Then the marginal distribution of B is given by

$$B \sim \mathcal{T}_{k,l}(v-k+1, \bar{B}, vS, N^{-1}). \tag{C.1}$$

Proof of Proposition 3. From the definition of the matrix normal-Wishart distribution given in Appendix B, we know that the joint density of B and H is

$$p(B,H) = \frac{|N|^{0.5k}\left(\frac{v}{2}\right)^{0.5kv}|S|^{0.5v}}{(2\pi)^{0.5kl}\,\Gamma_k\left(\frac{v}{2}\right)}|H|^{0.5(v+l-k-1)}$$
$$exp\{-0.5\,\mathrm{tr}[((B-\bar{B})N(B-\bar{B})'+vS)H]\}.$$

To ease visualization, define $\Psi := [(B-\bar{B})N(B-\bar{B})'+vS]^{-1}$. The marginal distribution of B is obtained by integrating out H:

$$p(B) = \int p(B,H)dH =$$
$$= \int \frac{|N|^{0.5k}\left(\frac{v}{2}\right)^{0.5kv}|S|^{0.5v}}{(2\pi)^{0.5kl}\,\Gamma_k\left(\frac{v}{2}\right)}|H|^{0.5(v+l-k-1)}exp[-0.5\,\mathrm{tr}(\Psi^{-1}H)]dH. \tag{C.2}$$

Now the trick is to multiply and divide this expression by

$$2^{0.5k(v+l)}\Gamma_k\left(\frac{v+l}{2}\right)|\Psi|^{0.5(v+l)}.$$

Then, by rearranging terms adequately, (C.2) becomes

$$p(B) = \frac{|N|^{0.5k}\left(\frac{v}{2}\right)^{0.5kv}|S|^{0.5v}}{(2\pi)^{0.5kl}\ \Gamma_k\left(\frac{v}{2}\right)} 2^{0.5k(v+l)}\Gamma_k\left(\frac{v+l}{2}\right)|\Psi|^{0.5(v+l)}$$
$$\int\left[2^{0.5k(v+l)}\Gamma_k\left(\frac{v+l}{2}\right)|\Psi|^{0.5(v+l)}\right]^{-1}$$
$$|H|^{0.5(v+l-k-1)}exp[-0.5\,\mathrm{tr}(\Psi^{-1}H)]dH. \tag{C.3}$$

Note that the integrand above is the PDF of a Wishart distribution for H with parameters $v+l$ and Ψ. Since it integrates to unity, we are left with

$$p(B) = \frac{|N|^{0.5k}\left(\frac{v}{2}\right)^{0.5kv}|S|^{0.5v}}{(2\pi)^{0.5kl}\ \Gamma_k\left(\frac{v}{2}\right)} 2^{0.5k(v+l)}\Gamma_k\left(\frac{v+l}{2}\right)|\Psi|^{0.5(v+l)}.$$

Now substitute Ψ and factor out vS from the determinant. Rearranging terms yields:

$$p(B) = \pi^{-0.5kl}\frac{\Gamma_k\left(\frac{v+l}{2}\right)}{\Gamma_k\left(\frac{v}{2}\right)}|N^{-1}|^{-0.5k}|vS|^{-0.5l}$$
$$|I_k + (vS)^{-1}(B-\bar{B})N(B-\bar{B})'|^{-0.5(v+l)}. \tag{C.4}$$

Note that $v+l = (v-k+1)+k+l-1$ and $v = (v-k+1)+k-1$. Then it is easy to see that (C.4) is the PDF of a matrix variate t distribution for B with parameters $v-k+1$, $\bar{B}$, vS and N^{-1}. Therefore, we conclude that

$$B \sim \mathcal{T}_{k,l}(v-k+1, \bar{B}, vS, N^{-1}).$$

□

Additionally, we present an adapted version of Theorem 2 of Uhlig (1997) (its proof can be found in appendix B of that paper).

Theorem 1 (Theorem 2 of Uhlig (1997))**.**
Given $v > k-1$ and $\lambda > 0$, let a prior for the $k \times l$ coefficient matrix B

and the $k(k+1)/2$ distinct elements of the precision matrix H be given by a density proportional to

$$f^{k,l}_{\mathcal{NW}}(B,H;\bar{B},N,v+1,S).$$

Suppose additionally that there is an unobserved shock to the precision matrix obeying

$$\tilde{H} = \mathcal{U}(H)'\Theta\mathcal{U}(H)/\lambda, \qquad \Theta \sim \mathcal{B}_k((v+l)/2, 1/2).$$

Then the posterior density for B and $\tilde{H}$ is proportional to

$$f^{k,l}_{\mathcal{NW}}(B,\tilde{H};\bar{B},\tilde{N},v,\tilde{S}),$$

where:

$$\tilde{N} = \lambda N,$$
$$\tilde{S} = \lambda \frac{v+1}{v} S.$$

This result is used in the proof of our Proposition 2 below.

Proof of Proposition 1. In period t, the updated distribution is obtained through Bayes' rule (RISTIC et al., 2004):

$$p(B_t,H_t|Y_t) = \frac{p(y_t|B_t,H_t)p(B_t,H_t|Y_{t-1})}{p(y_t|Y_{t-1})} \propto p(y_t|B_t,H_t)p(B_t,H_t|Y_{t-1}).$$

From the measurement equation, (2.3), we have that

$$y_t|B_t,H_t \sim \mathcal{N}(B_tX_t, H_t^{-1}).$$

And by assumption, we have the initial condition that

$$B_t,H_t|Y_{t-1} \sim \mathcal{NW}_{k,l}(B_{t|t-1}, N_{t|t-1}, v, S_{t|t-1}),$$

with all parameters know and $N_{t|t-1}$ symmetric. Multiplying the PDFs of these two distributions yields:

$$p(B_t,H_t|Y_t) \propto p(y_t|B_t,H_t)p(B_t,H_t|Y_{t-1}) =$$

$$= (2\pi)^{-0.5k}|H_t^{-1}|^{-1}|N_{t|t-1}|^{0.5k}(v/2)^{0.5kv}|S_{t|t-1}|^{0.5v}(2\pi)^{-0.5kl}$$
$$\Gamma_k(v/2)^{-1}|H_t|^{0.5(l+v-k-1)}exp\{-0.5\,\mathrm{tr}[(y_t - B_tX_t)(y_t - B_tX_t)' \cdots$$
$$+ (B_t - B_{t|t-1})N_{t|t-1}(B_t - B_{t|t-1})' + vS_{t|t-1}]H_t\}.$$

Defining $N_{t|t} := N_{t|t-1} + X_tX_t'$, it is possible to write the terms within square brackets as:

$$B_tN_{t|t}B_t' - B_tN_{t|t}(N_{t|t}^{-1})'(B_{t|t-1}N_{t|t-1} + y_tX_t')'$$
$$- (B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1}N_{t|t}B_t'$$
$$+ y_ty_t' + B_{t|t-1}N_{t|t-1}B_{t|t-1}' + vS_{t|t-1}.$$

Note that one can complete the square in this expression by adding and subtracting

$$(B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1}N_{t|t}(N_{t|t}^{-1})'(B_{t|t-1}N_{t|t-1} + y_tX_t')'.$$

This yields:

$$[B_t - (B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1}]N_{t|t}[B_t - (B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1}]'$$
$$+ y_ty_t' + B_{t|t-1}N_{t|t-1}B_{t|t-1}' + vS_{t|t-1}$$
$$- (B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1}N_{t|t}(N_{t|t}^{-1})'(B_{t|t-1}N_{t|t-1} + y_tX_t')'.$$

Defining

$$B_{t|t} = (B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1},$$
$$e_t = y_t - B_{t|t-1}X_t, \text{ and}$$
$$S_{t|t} = \frac{v}{v+1}S_{t|t-1} + \frac{1}{v+1}e_t(1 - X_t'N_{t|t}^{-1}X_t)e_t',$$

it is possible to show that

$$y_ty_t' + B_{t|t-1}N_{t|t-1}B_{t|t-1}' + vS_{t|t-1}$$
$$- (B_{t|t-1}N_{t|t-1} + y_tX_t')N_{t|t}^{-1}N_{t|t}(N_{t|t}^{-1})'(B_{t|t-1}N_{t|t-1} + y_tX_t')' =$$
$$= (v+1)S_{t|t}.$$

So one can write the updated density above as

$$p(B_t, H_t|Y_t) \propto (2\pi)^{-0.5k(l+1)}|N_{t|t-1}|^{0.5k}(v/2)^{0.5kv}$$

$$|S_{t|t-1}|^{0.5v}\Gamma_k(v/2)^{-1}|H_t|^{0.5(l+v+1-k-1)}$$
$$exp\{-0.5\,\mathrm{tr}[((B_t - B_{t|t})N_{t|t}(B_t - B_{t|t})' + (v+1)S_{t|t})H_t]\}. \tag{C.5}$$

The second line of this expression is the kernel of a normal-Wishart density for B_t and H_t conditional on Y_t, with parameters $B_{t|t}$, $N_{t|t}$, $v+1$ and $S_{t|t}$. Therefore, we can conclude that

$$B_t, H_t|Y_t \sim \mathcal{NW}_{k,l}(B_{t|t}, N_{t|t}, v+1, S_{t|t}),$$

where:

$$B_{t|t} = (B_{t|t-1}N_{t|t-1} + y_t X_t')N_{t|t}^{-1},$$
$$N_{t|t} = N_{t|t-1} + X_t X_t',$$
$$S_{t|t} = \frac{v}{v+1}S_{t|t-1} + \frac{1}{v+1}e_t(1 - X_t' N_{t|t}^{-1} X_t)e_t',$$
$$e_t = y_t - B_{t|t-1}X_t.$$

This proves the claim about the filtered joint density. From the definition of the matrix normal-Wishart distribution, it follows directly that

$$H_t|Y_t \sim \mathcal{W}_{k,l}[v+1, (v+1)^{-1}S_{t|t}^{-1}],$$

and

$$B_t|H_t, Y_t \sim \mathcal{N}_{k,l}(B_{t|t}, H_t^{-1} \otimes N_{t|t}^{-1}).$$

Then it follows from Proposition 3 above that the marginal filtered density of the coefficients is

$$B_t|Y_t \sim \mathcal{T}_{k,l}[v-k+2, B_{t|t}, (v+1)S_{t|t}, N_{t|t}^{-1}].$$

□

***Proof of Proposition** 2.* In period $t+1$, the predictive density based on information available up to period t is obtained via the Chapman-Kolmogorov equation (RISTIC et al., 2004):

$$p(B_{t+1}, H_{t+1}|Y_t) = \int\int p(B_{t+1}, H_{t+1}|B_t, H_t)p(B_t, H_t|Y_t)dH_t dB_t$$

$$= \int \int p(B_{t+1}|B_t, H_t, H_{t+1}) p(H_{t+1}|B_t, H_t) p(B_t, H_t|Y_t) dH_t dB_t.$$

But the transition equations, (2.2) and (2.6), imply that

$$p(B_{t+1}|B_t, H_t, H_{t+1}) = p(B_{t+1}|B_t, H_{t+1}) \text{ and}$$
$$p(H_{t+1}|B_t, H_t) = p(H_{t+1}|H_t).$$

So one can write

$$p(B_{t+1}, H_{t+1}|Y_t) = \int p(B_{t+1}|B_t, H_{t+1}) \int p(H_{t+1}|H_t) p(B_t, H_t|Y_t) dH_t dB_t.$$

The integral with respect to H_t can be solved using Theorem 2 of Uhlig (1997) (Theorem 1 above). It yields:

$$\int p(H_{t+1}|H_t) p(B_t, H_t|Y_t) dH_t = f^{k,l}_{\mathcal{NW}}(B_t, H_{t+1}|Y_t; B_{t|t}, \lambda N_{t|t}, v, S_{t+1|t}), \tag{C.6}$$

where $S_{t+1|t} = \lambda \frac{(v+1)}{v} S_{t|t}$, while $B_{t|t}$, $N_{t|t}$, v and $S_{t|t}$ are assumed to be known. So the predictive density above becomes

$$\begin{aligned} p(B_{t+1}, H_{t+1}|Y_t) &= \int p(B_{t+1}|B_t, H_{t+1}) p(B_t, H_{t+1}|Y_t) dB_t \\ &= \int p(B_{t+1}|B_t, H_{t+1}) p(B_t|H_{t+1}, Y_t) p(H_{t+1}|Y_t) dB_t \\ &= p(H_{t+1}|Y_t) \int p(B_{t+1}|B_t, H_{t+1}) p(B_t|H_{t+1}, Y_t) dB_t. \end{aligned} \tag{C.7}$$

From (2.6), we know $p(B_{t+1}|B_t, H_{t+1})$. And from (C.6) above plus the definition of the matrix normal-Wishart distribution, we obtain $p(B_t|H_{t+1}, Y_t)$. It is then straightforward to combine these two densities to solve the integral with respect to B_t:

$$\int (2\pi)^{-0.5kl} |H^{-1}_{t+1}|^{-0.5l} |Q^{-1}|^{-0.5k}$$
$$exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_{t+1} - B_t) Q (B_{t+1} - B_t)']\}$$
$$(2\pi)^{-0.5kl} |H^{-1}_{t+1}|^{-0.5l} |(\lambda N_{t|t})^{-1}|^{-0.5k}$$

$$exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_t - B_{t|t})\lambda N_{t|t}(B_t - B_{t|t})']\}dB_t =$$
$$= (2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|Q^{-1}|^{-0.5k}(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}$$
$$|(\lambda N_{t|t})^{-1}|^{-0.5k}\int exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_t - B_{t+1})Q(B_t - B_{t+1})'\cdots$$
$$+ H_{t+1}(B_t - B_{t|t})\lambda N_{t|t}(B_t - B_{t|t})']\}dB_t. \tag{C.8}$$

The expression within square brackets can be rearranged as

$$H_{t+1}[B_t(Q + \lambda N_{t|t})B_t' - B_t(QB_{t+1} + \lambda N_{t|t}B_{t|t})'$$
$$-(B_{t+1}Q + B_{t|t}\lambda N_{t|t})B_t' + B_{t+1}QB_{t+1}' + B_{t|t}\lambda N_{t|t}B_{t|t}'].$$

Now to complete the square with respect to B_t, we add and subtract

$$(B_{t+1}Q + B_{t|t}\lambda N_{t|t})(Q + \lambda N_{t|t})^{-1}(B_{t+1}Q + B_{t|t}\lambda N_{t|t})'.$$

This yields:

$$H_{t+1}(B_t - \mu)\Omega(B_t - \mu)' + H_{t+1}(B_{t+1}QB_{t+1}' + B_{t|t}\lambda N_{t|t}B_{t|t}' - \mu\Omega\mu'),$$

where $\mu := (B_{t+1}Q + B_{t|t}\lambda N_{t|t})(Q + \lambda N_{t|t})^{-1}$ and $\Omega := Q + \lambda N_{t|t}$. To ease visualization, define $C := H_{t+1}(B_{t+1}QB_{t+1}' + B_{t|t}\lambda N_{t|t}B_{t|t}' - \mu\Omega\mu')$. Then (C.8) becomes:

$$(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|Q^{-1}|^{-0.5k}(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|(\lambda N_{t|t})^{-1}|^{-\frac{k}{2}}$$
$$\int exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_t - \mu)\Omega(B_t - \mu)' + C]\}dB_t =$$
$$= (2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|Q^{-1}|^{-0.5k}(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|(\lambda N_{t|t})^{-1}|^{-\frac{k}{2}}$$
$$exp\{-0.5\,\mathrm{tr}[C]\}\int exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_t - \mu)\Omega(B_t - \mu)']\}dB_t. \tag{C.9}$$

Note that the integrand above is the kernel of a matrix variate normal distribution for B_t with mean matrix μ and covariance matrix $(H_{t+1} \otimes \Omega)^{-1}$. So we can solve the integral using the fact that

$$\int K(B_t)dB_t = \frac{K(B_t)}{D(B_t)} = \kappa(B_t)^{-1},$$

where $D(.)$ is the PDF, $K(.)$ is the kernel and $\kappa(.)$ is the constant of integration of the density for B_t (BAUWENS et al., 1999). This yields:

$$\int exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_t - \mu)\Omega(B_t - \mu)']\}dB_t =$$

$$= [(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|\Omega^{-1}|^{-0.5k}]^{-1}.$$

Plugging this into (C.9) and the resulting expression back into (C.7) yields:

$$p(B_{t+1}, H_{t+1}|Y_t) = p(H_{t+1}|Y_t)(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|Q^{-1}|^{-0.5k}$$
$$|(\lambda N_{t|t})^{-1}|^{-0.5k}|\Omega^{-1}|^{-0.5k} exp\{-0.5\,\mathrm{tr}[C]\}. \quad \text{(C.10)}$$

Now recall that $C := H_{t+1}(B_{t+1}QB'_{t+1} + B_{t|t}\lambda N_{t|t}B'_{t|t} - \mu\Omega\mu')$. It is possible to simplify this by plugging in μ and Ω and by using the following matrix identities[1]:

$$(A+CBC')^{-1} = A^{-1} - A^{-1}C(B^{-1}+C'A^{-1}C)^{-1}C'A^{-1},$$
$$(A^{-1}+B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A.$$

Then C can be written as:

$$C = H_{t+1}(B_{t+1} - B_{t|t})(Q^{-1} + \lambda^{-1}N_{t|t}^{-1})^{-1}(B_{t+1} - B_{t|t})'.$$

Additionally, note that

$$|Q^{-1}|^{-0.5k}|(\lambda N_{t|t})^{-1}|^{-0.5k}|\Omega^{-1}|^{-0.5k} = [|Q||(Q+\lambda N_{t|t})^{-1}||\lambda N_{t|t}|]^{0.5k}$$
$$= [|Q(Q+\lambda N_{t|t})^{-1}\lambda N_{t|t}|]^{0.5k}$$
$$= [|QQ^{-1}(Q^{-1}+\lambda^{-1}N_{t|t}^{-1})^{-1}\lambda^{-1}N_{t|t}^{-1}\lambda N_{t|t}|]^{0.5k}$$
$$= [|Q^{-1}+\lambda^{-1}N_{t|t}^{-1}|]^{-0.5k}.$$

Then (C.10) becomes:

$$p(B_{t+1}, H_{t+1}|Y_t) =$$
$$= p(H_{t+1}|Y_t)(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|Q^{-1} + \lambda^{-1}N_{t|t}^{-1}|^{-0.5k}$$
$$exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_{t+1} - B_{t|t})(Q^{-1} + \lambda^{-1}N_{t|t}^{-1})^{-1}(B_{t+1} - B_{t|t})']\}. \quad \text{(C.11)}$$

It is clear that

$$(2\pi)^{-0.5kl}|H_{t+1}^{-1}|^{-0.5l}|Q^{-1} + \lambda^{-1}N_{t|t}^{-1}|^{-0.5k}$$

[1] See section 3.2 of Petersen & Pedersen (2012).

$$exp\{-0.5\,\mathrm{tr}[H_{t+1}(B_{t+1}-B_{t|t})(Q^{-1}+\lambda^{-1}N_{t|t}^{-1})^{-1}(B_{t+1}-B_{t|t})']\}$$

is the PDF of a matrix Normal distribution for $B_{t+1}|H_{t+1}, Y_t$ with mean matrix $B_{t|t}$ and covariance matrix $H_{t+1}^{-1} \otimes (Q^{-1}+\lambda^{-1}N_{t|t}^{-1})$. Additionally, we know from (C.6) that

$$H_{t+1}|Y_t \sim \mathcal{W}_{k,l}[v,(vS_{t+1|1})^{-1}].$$

Then it follows from the definition of the Normal-Wishart distribution that:

$$B_{t+1}, H_{t+1}|Y_t \sim \mathcal{NW}_{k,l}(B_{t|t}, N_{t+1|t}, v, S_{t+1|t}), \tag{C.12}$$

where:

$$\begin{aligned} B_{t+1|t} &= B_{t|t}, \\ N_{t+1|t} &= [Q^{-1}+(\lambda N_{t|t})^{-1}]^{-1}, \\ S_{t+1|t} &= \lambda\frac{(v+1)}{v}S_{t|t}. \end{aligned}$$

This proves the first part of the theorem. From the definition of the matrix Normal-Wishart distribution, it follows directly that

$$H_{t+1}|Y_t \sim \mathcal{W}_{k,l}(v, v^{-1}S_{t+1|t}^{-1}),$$

and

$$B_{t+1}|H_{t+1}, Y_t \sim \mathcal{N}_{k,l}(B_{t+1|t}, H_{t+1}^{-1} \otimes N_{t+1|t}^{-1}).$$

Then once again from Proposition 3 above we have that the marginal predictive density of the coefficients is

$$B_{t+1}|Y_t \sim \mathcal{T}_{k,l}(v-k+1, B_{t+1|t}, vS_{t+1|t}, N_{t+1|t}^{-1}).$$

□

Proof of Corollary 1. As seen from (C.5) in the proof of Proposition 1, the updated joint density of the states can be written as:

$p(B_t, H_t|Y_t) \propto$

$$(2\pi)^{-0.5k(l+1)}|N_{t|t-1}|^{0.5k}(v/2)^{0.5kv}|S_{t|t-1}|^{0.5v}\Gamma_k(v/2)^{-1}|H_t|^{0.5(l+v-k)}$$
$$exp\{-0.5\,\mathrm{tr}[((B_t-B_{t|t})N_{t|t}(B_t-B_{t|t})'+(v+1)S_{t|t})H_t]\}.$$

We know that the second line of this expression is the kernel of a Normal-Wishart density for B_t and H_t conditional on Y_t, with parameters $B_{t|t}$, $N_{t|t}$, $v+1$ and $S_{t|t}$. This implies that

$$B_t, H_t|Y_t \sim \mathcal{NW}_{k,l}(B_{t|t}, N_{t|t}, v+1, S_{t|t}),$$

where:

$$\begin{aligned}
B_{t|t} &= (B_{t|t-1}N_{t|t-1} + y_t X_t')N_{t|t}^{-1},\\
N_{t|t} &= N_{t|t-1} + X_t X_t',\\
S_{t|t} &= \frac{v}{v+1}S_{t|t-1} + \frac{1}{v+1}e_t(1 - X_t' N_{t|t}^{-1} X_t)e_t',\\
e_t &= y_t - B_{t|t-1}X_t.
\end{aligned}$$

The proportionality symbol in (C.5) is due to the fact that we are omitting the normalizing constant in the denominator, $p(y_t|Y_{t-1})$. The terms

$$(2\pi)^{-0.5k(l+1)}|N_{t|t-1}|^{0.5k}(v/2)^{0.5kv}|S_{t|t-1}|^{0.5v}\Gamma_k(v/2)^{-1}$$

do not enter the kernel of the density and are regarded as constants. We know that these terms divided by $p(y_t|Y_{t-1})$ will equal the constant of integration of $f^{k,l}_{\mathcal{NW}}(B_t, H_t|Y_t; B_{t|t}, N_{t|t}, v+1, S_{t|t})$. I.e.,

$$\frac{(2\pi)^{-0.5k(l+1)}|N_{t|t-1}|^{0.5k}(v/2)^{0.5kv}|S_{t|t-1}|^{0.5v}\Gamma_k(v/2)^{-1}}{p(y_t|Y_{t-1})} =$$
$$= \frac{|N_{t|t}|^{0.5k}((v+1)/2)^{0.5k(v+1)}|S_{t|t}|^{0.5(v+1)}}{(2\pi)^{0.5lk}\Gamma_k((v+1)/2)}.$$

See κ_{NW} in Appendix B. Now plug $S_{t|t}$ and e_t into this and solve for $p(y_t|Y_{t-1})$:

$$p(y_t|Y_{t-1}) =$$
$$\pi^{-0.5k}\frac{\Gamma_k[(v+1)/2]}{\Gamma_k(v/2)}\frac{v^{0.5kv}}{(v+1)^{0.5k(v+1)}}|N_{t|t-1}|^{0.5k}|N_{t|t}|^{-0.5k}|S_{t|t-1}|^{0.5v}$$

$$\left| \frac{v}{v+1} S_{t|t-1} + \frac{1}{v+1} (y_t - B_{t|t-1} X_t)(1 - X_t' N_{t|t}^{-1} X_t)(y_t - B_{t|t-1} X_t)' \right|^{\frac{v+1}{-2}}.$$

Using the property that $|AB| = |A||B|$, we can factor out $v/(v+1)S_{t|t-1}$. And noting that $(1 - X_t' N_{t|t}^{-1} X_t)$ is a scalar, we can rearrange the expression above to obtain

$$p(y_t|Y_{t-1}) =$$
$$\pi^{-0.5k} \frac{\Gamma_k[(v+1)/2]}{\Gamma_k(v/2)} \frac{v^{0.5kv}}{(v+1)^{0.5k(v+1)}} |N_{t|t-1}|^{\frac{k}{2}} |N_{t|t}|^{\frac{-k}{2}} |S_{t|t-1}|^{\frac{-1}{2}} v^{\frac{-k}{2}}$$
$$|I_k + (1 - X_t' N_{t|t}^{-1} X_t)(v S_{t|t-1})^{-1} (y_t - B_{t|t-1} X_t)(y_t - B_{t|t-1} X_t)'|^{-\frac{v+1}{2}}. \tag{C.13}$$

Now define

$$\Sigma_t := (1 - X_t' N_{t|t}^{-1} X_t)(v S_{t|t-1})^{-1}. \tag{C.14}$$

It is possible to show that

$$|N_{t|t-1}|^{0.5k} |N_{t|t}|^{-0.5k} |S_{t|t-1}|^{-0.5} v^{-0.5k} = |\Sigma_t|^{-0.5}. \tag{C.15}$$

From Theorem 1.4.1 of Gupta & Nagar (2000), it follows that

$$\frac{\Gamma_k[(v+1)/2]}{\Gamma_k(v/2)} = \frac{\Gamma[(v+1)/2]}{\Gamma(v+1-k/2)}. \tag{C.16}$$

Additionally, it is easy to use the Sylvester's determinant identity to verify that

$$|I_k + \Sigma_t (y_t - B_{t|t-1} X_t)(y_t - B_{t|t-1} X_t)'|^{-0.5(v+1)} =$$
$$= 1 + (y_t - B_{t|t-1} X_t)' \Sigma_t (y_t - B_{t|t-1} X_t). \tag{C.17}$$

Note that here we used the assumption that $S_{t|t-1}$ is symmetric. Then (C.13) becomes

$$p(y_t|Y_{t-1}) =$$
$$= \pi^{-\frac{k}{2}} \frac{\Gamma[(v+1)/2]}{\Gamma(v+1-k/2)} |\Sigma_t|^{\frac{1}{2}} [1 + (y_t - B_{t|t-1} X_t)' \Sigma_t (y_t - B_{t|t-1} X_t)]^{-\frac{v+1}{2}}$$
$$= \pi^{-0.5k} \frac{\Gamma[(v+1)/2]}{\Gamma(v+1-k/2)} |\Sigma_t|^{0.5} (v+1-k)^{0.5k} (v+1-k)^{-0.5k}$$

$$[1+(y_t - B_{t|t-1}X_t)'(v+1-k)^{-1}(v+1-k)\Sigma_t(y_t - B_{t|t-1}X_t)]^{-\frac{v+1}{2}}$$
$$= [(v+1-k)\pi]^{-0.5k}\frac{\Gamma[(v+1-k+k)/2]}{\Gamma(v+1-k/2)}|(v+1-k)^{-1}\Sigma_t^{-1}|^{-0.5}$$
$$\{1+(v+1-k)^{-1}(y_t - B_{t|t-1}X_t)'[(v+1-k)^{-1}\Sigma_t^{-1}]^{-1}\cdots$$
$$(y_t - B_{t|t-1}X_t)\}^{-\frac{v+1-k+k}{2}}. \tag{C.18}$$

Note that this is the PDF of a multivariate t distribution with $v+1-k$ degrees of freedom, location parameter $B_{t|t-1}X_t$ and scale matrix $(v+1-k)^{-1}\Sigma_t^{-1}$ [see Appendix A.1.17 of Greenberg (2008)]. Therefore, we conclude that

$$y_t|Y_{t-1} \sim t_k[v+1-k, B_{t|t-1}X_t, (v+1-k)^{-1}\Sigma_t^{-1}], \tag{C.19}$$

where

$$\Sigma_t = (1 - X_t' N_{t|t}^{-1} X_t)(vS_{t|t-1})^{-1},$$
$$N_{t|t} = N_{t|t-1} + X_t X_t'.$$

□