Large Vector Autoregressions with asymmetric priors and time varying volatilities*

Andrea Carriero
Queen Mary, University of London
a.carriero@qmul.ac.uk

Todd E. Clark
Federal Reserve Bank of Cleveland
todd.clark@clev.frb.org

Massimiliano Marcellino
Bocconi University, IGIER and CEPR
massimiliano.marcellino@unibocconi.it

This draft: November 2015

Abstract

We propose a new algorithm which allows easy estimation of Vector Autoregressions (VARs) featuring asymmetric priors and time varying volatilities, even when the cross sectional dimension of the system N is particularly large. The algorithm is based on a simple triangularisation which allows to simulate the conditional mean coefficients of the VAR by drawing them equation by equation. This strategy reduces the computational complexity by a factor of N^2 with respect to the existing algorithms routinely used in the literature and by practitioners. Importantly, this new algorithm can be easily obtained by modifying just one of the steps of the existing algorithms. We illustrate the benefits of the algorithm with numerical and empirical applications.

Keywords: Bayesian VARs, stochastic volatility, large datasets, forecasting, impulse response functions.

J.E.L. Classification: C11, C13, C33, C53.

*The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Cleveland or the Federal Reserve System. Carriero gratefully acknowledges support for this work from the Economic and Social Research Council [ES/K010611/1].

1 Introduction

The recent literature has shown that two main ingredients are key for the specification of a good Vector Autoregressive model (VAR) for forecasting and structural analysis of macroeconomic data: a large cross section of macroeconomic variables, and modeling time variation in their volatilities. Contributions which highlighted the importance of using a large information set include Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015) and Koop (2013), which all point out that large systems perform better than smaller systems in forecasting and structural analysis. Contributions that have highlighted the importance of time variation in the volatilities include Clark (2011), Clark and Ravazzolo (2015), Cogley and Sargent (2005), D'Agostino, Gambetti and Giannone (2013), and Primiceri (2005).

Even though it is now clear that it would be ideal to include both of these features when specifying a VAR model for macroeconomic variables, there are no papers which jointly allow for time variation and large datasets. To the best of our knowledge, the only two exceptions are Koop and Korobilis (2013) and Carriero, Clark, and Marcellino (2012). Koop and Korobilis (2013) propose a computational (not fully Bayesian) shortcut that allows for time-varying volatility, roughly speaking, using a form of exponential smoothing of volatility that allows them to estimate a large VAR. However, the resulting estimates are not fully Bayesian and do not allow, for example, to compute the uncertainty around the volatility estimates in a coherent fashion. Our previous work in Carriero, Clark, and Marcellino (2012) also tries to tackle this issue, by assuming a specific structure for the volatilities in the VAR. In particular, in a common stochastic volatility specification, we imposed a factor structure on the volatilities and further assumed that i) there is no idiosyncratic component for the conditional volatilities, and ii) all the conditional volatilities have a factor loading of 1, which implies that the order of magnitude of the movements in volatility is proportional across variables. Although the evidence in Carriero, Clark, and Marcellino (2012) indicates that the proposed model improves over an homoskedastic VAR in density forecasting, the restrictions discussed above do not necessarily hold in a typical dataset of macroeconomic and financial variables, especially so as the cross-sectional dimension grows. Some researchers might prefer not to impose the restrictions, out of concern for misspecification.

The reason why stochastic volatilities in the disturbance term can not easily be estimated in a large VAR — without restrictions such as those of Carriero, Clark, and Marcellino (2012) — lies in the structure of the likelihood function. The introduction of drifting volatilities leads to the loss of symmetry in the model, which in turn implies that estimation of the system becomes rapidly unmanageable. Homoskedastic VAR models are SUR mod-

els featuring the same set of regressors in each equation. This symmetry across equations means that homoskedastic VAR models have a Kronecker structure in the likelihood, and can therefore be estimated via OLS equation by equation. In a Bayesian setting the symmetry in the likelihood transfers to the posterior, as long as the prior used also features a Kronecker structure. Equation-specific stochastic volatility breaks this symmetry because each equation is driven by a different volatility. This implies that the model needs to be vectorised before estimation. The challenge with such a model is that drawing the VAR coefficients from the conditional posterior involves computing a (variance) matrix with the number of rows and columns equal to the number of variables squared times the number of lags (plus one if a constant is included). The size of this matrix increases with the square of the number of variables in the model, making CPU time requirements highly nonlinear in the number of variables.

Similarly, there are cases in which even in presence of a symmetric likelihood function, the prior distribution on the coefficients is not symmetric and this again implies a considerable increase in the computational complexity of the model. For example, the VAR estimated by Banbura, Giannone, and Reichlin (2010) is a homoskedastic VAR with 130 variables, but in order to make this estimation possible a specific structure must be assumed for the prior distribution of the coefficients. In particular, the original Litterman (1986) implementation of the so called Minnesota prior puts additional shrinkage on the lags of all the variables other than the dependent variable of the i-th VAR equation, in order to capture the idea that, at least in principle, these lags should be less relevant than the lag of the dependent variable itself. But such kind of shrinkage can not be implemented in the model of Banbura, Giannone, and Reichlin (2010) without losing the Kronecker structure of the prior. In this case the prior is not symmetric across equations and therefore, even in presence of a symmetric likelihood, the resulting posterior is not symmetric across equations, which implies that the system needs to be vectorised prior to estimation, which in turn results in the same type of computational costs we described in the previous paragraph. Incidentally, it is for this reason that Litterman (1986) assumed a (fixed) diagonal prior variance for the disturbance term, since this assumption allows to estimate his model equation by equation.

To summarize, if either the prior or the likelihood induce an asymmetry in the posterior of the VAR coefficients, the model needs to be vectorised and its computational complexity rises from N^3 up to N^6, where N is the size of the cross section. For this reason the only VAR which can be reasonably estimated with a large cross section of data is the one proposed by Kadiyala and Karlsson (1997), which features symmetry in both the prior and the likelihood, and it is indeed on this model that papers such as Banbura, Giannone, and

Reichlin (2010) and Carriero, Clark, and Marcellino (2012) are built on.

In this paper we propose a new algorithm which allows to estimate VARs featuring asymmetries either in the prior or in the likelihood, thereby allowing for models with asymmetric priors and time varying volatilities. The new algorithm is based on a simple triangularisation of the VAR, which allows to simulate the VAR coefficients by drawing them equation by equation. The new algorithm is very simple and, importantly, it can be easily inserted in any pre-existing algorithm for estimation of VAR models. This algorithm reduces the computational complexity for estimating the VAR model to the order N^4, which is considerably faster than the complexity N^6 arising from the traditional algorithm, and therefore it allows to estimate large models.

The paper is structured as follows. In Section 2 we present the model and the estimation algorithm. Section 3 presents a numerical comparison to illustrate the gains in terms of computing time (and convergence and mixing properties). Section 4 discusses an empirical application where we compute responses to a monetary policy shock in a large VAR with time varying volatilities. Section 5 concludes.

2 An estimation algorithm for large BVARs

2.1 The model

Consider the following VAR model with stochastic volatility:

$$y_t = \Pi_0 + \Pi(L)y_{t-1} + v_t, \tag{1}$$

$$v_t = A^{-1}\Lambda_t^{0.5}\epsilon_t,\ \epsilon_t \sim iid\ N(0, I_T), \tag{2}$$

where $t = 1, ..., T$, the dimension of the vectors y_t, v_t and ϵ_t is N, Λ_t is a diagonal matrix with generic j-th element $h_{j,t}$ and A^{-1} is a lower triangular matrix with ones on its main diagonal. The specification above implies a time varying variance for the disturbances v_t:

$$\Sigma_t \equiv Var(v_t) = A^{-1}\Lambda_t A^{-1\prime}. \tag{3}$$

The diagonality of the matrix Λ_t implies that the generic j-th element of the rescaled VAR disturbances $\tilde{v}_t = Av_t$ is given by $\tilde{v}_{j,t} = h_{j,t}^{0.5}\epsilon_{jt}$. Taking logs of squares of $\tilde{v}_{j,t}$ yields the following set of observation equations:

$$\ln \tilde{v}_{j,t}^2 = \ln h_{j,t} + \ln \epsilon_{j,t}^2,\ j = 1, \ldots, N. \tag{4}$$

The model is completed by specifying laws of motion for the unobserved states:

$$\ln h_{j,t} = \ln h_{j,t-1} + e_{j,t},\ j = 1, \ldots, N, \tag{5}$$

where the vector of innovations to volatilities e_t is $N(0, \Phi)$ (and independent across time), with a variance matrix Φ that is full matrix as in Primiceri (2005) and not diagonal as in Cogley and Sargent (2005).

In equation (2) we do not allow the elements in A^{-1} to vary over time, which would yield the variance specification of Primiceri (2005). We do so because Primiceri (2005) found little variation in such coefficients, and specifying variation in these coefficients would imply additional $N(N-1)/2$ state equations such as (5). Note however that even if one were to specify also A^{-1} as time varying, this would not impact the main computational advantage arising from the algorithm we will propose below, as the main bottleneck in estimating large VARs is the inversion of the variance matrix of the $\Pi(L)$ coefficients, not the simulation of the drifting covariances and volatilities. Similarly, one can modify equation (5) so that the states $\ln h_{j,t}$ follow an autoregressive process rather than a random walk, but again this is not essential to the point we make in this paper.

In a Bayesian setting, to estimate the model the likelihood needs to be combined with a prior distribution for the model coefficients

$$\Theta = \{\Pi, A, \Phi\} \tag{6}$$

and the unobserved states Λ_t. Under the conventional systems approach, the priors for the coefficients blocks of the model are as follows:

$$\begin{aligned} \text{vec}(\Pi) &\sim N(\text{vec}(\underline{\mu}_\Pi), \underline{\Omega}_\Pi); & (7)\\ A &\sim N(\underline{\mu}_A, \underline{\Omega}_A); & (8)\\ \Phi &\sim IW(\underline{d}_\Phi \cdot \underline{\Phi}, \underline{d}_\Phi). & (9) \end{aligned}$$

The model is completed by eliciting a prior for the initial value of the state variables Λ_t which we set as diffuse.

2.2 Model estimation

The model presented above is typically estimated as follows. First, the conditional posterior distributions of all the coefficients blocks are derived:

$$\begin{aligned} \text{vec}(\Pi)|A, \Lambda_T, y_T &\sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi); & (10)\\ A|\Pi, \Lambda_T, y_T &\sim N(\bar{\mu}_A, \overline{\Omega}_A); & (11)\\ \Phi|\Lambda_T, y_T &\sim IW((\underline{d}_\Phi + T) \cdot \bar{\Phi}, \underline{d}_\Phi + T), & (12) \end{aligned}$$

where Λ_T and y_T denote the history of the states and data up to time T, and where the posterior moments $\bar{\mu}_\Pi$, $\overline{\Omega}_\Pi$, $\bar{\mu}_A$, $\overline{\Omega}_A$ and $\bar{\Phi}$ can be derived by combining prior moments and

likelihood moments.[1]

A step of a Gibbs sampler cycling through (10)-(12) provides a draw from the joint posterior distribution $p(\Theta|\Lambda_T, y_T)$. Conditional on this draw, a draw from the distribution of the states $p(\Lambda_T|\Theta, y_T)$ is obtained using the observation and transition equations (4) and (5), by using either the independent Metropolis algorithm proposed by Jacquier, Polson and Rossi (1994) or the mixture of normals approximation algorithm proposed by Kim, Shepard and Chib (1998).[2] Cycling through $p(\Theta|\Lambda_T, y_T)$ and $p(\Lambda_T|\Theta, y_T)$ provides the joint posterior of the model coefficients and unobserved states $p(\Theta, \Lambda_T|y_T)$. This estimation strategy is used in all of the implementations of this model.

In this paper we are interested in one specific step of the algorithm described above, the draw from $\Pi|A, \Lambda_T, y_T$ described in equation (10). The main problem in this step is that — as is clear from the fact that equation (10) is specified in terms of the vectorised vector of coefficients vec(Π) — it involves the manipulation of the variance matrix of the coefficients Π, which is a square matrix of dimension $N(Np+1)$.

Consider drawing $m = 1, ..., M$ draws from the posterior of Π. To perform a draw Π^m from (10), one needs to draw a $N(Np+1)-$dimensional random vector (distributed as a standard Gaussian), rand, and to compute:

$$\text{vec}(\Pi^m) = \bar{\Omega}_\Pi \left\{ \text{vec}\left(\sum_{t=1}^{T} X_t y_t' \Sigma_t^{-1} \right) + \underline{\Omega}_\Pi^{-1} \text{vec}(\underline{\mu}_\Pi) \right\} + chol(\bar{\Omega}_\Pi) \times \text{rand} \tag{13}$$

The calculation above involves computations of the order of $4O(N^6)$. Indeed, it is necessary to compute: i) the matrix $\bar{\Omega}_\Pi$ by inverting

$$\bar{\Omega}_\Pi^{-1} = \underline{\Omega}_\Pi^{-1} + \sum_{t=1}^{T} (\Sigma_t^{-1} \otimes X_t X_t'); \tag{14}$$

ii) its Cholesky factor $chol(\bar{\Omega}_\Pi)$; iii) multiply the matrices obtained in i) and ii) by the vector in the curly brackets of (13) and the vector rand respectively. Since each of these operations requires $O(N^6)$ elementary operations, the total computational complexity to compute a draw Π^m is $4 \times O(N^6)$. Also computation of $\underline{\Omega}_\Pi^{-1}\text{vec}(\underline{\mu}_\Pi)$ requires $O(N^6)$ operations but this is fixed across repetitions so it needs to be computed just once.[3]

[1] Note that knowledge on the full history of the states Λ_T renders redundant conditioning on the hyper-parameters Φ regulating the law of motions of such states when drawing Π and A, as well as conditioning on Π and A when drawing Φ.

[2] In such case one needs to introduce another set of state variables s_T used to approximate the error term appearing in (4). For more details see Section (2.5.1) below.

[3] Some speed improvements can be obtained as follows. Define $\bar{\Omega}_\Pi^{-1} = C'C$ where C is an upper triangular matrix and C' is therefore the Cholesky factor of $\bar{\Omega}_\Pi^{-1}$. It follows that $\bar{\Omega}_\Pi = C^{-1}C'^{-1}$ with C^{-1} upper

For a system of 20 variables, which is the "medium" size considered in studies such as Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2012), Giannone, Lenza, and Primiceri (2015) and Koop (2013) this amounts to $4 \times 20^6 = 256$ millions elementary operations (per single draw), and this is the main bottleneck that prevented the existing literature to estimate these models with more than a handful of variables, typically 3 to 5.

2.3 Asymmetric priors

It is important to note that the computational problem arises from the fact that in a stochastic volatility model, if we rescale each of the equations by the error volatility, in a weighted least squares fashion, then each equation ends up having different regressors, and this is the root of the asymmetry in the likelihood. However, the computational problem of the dimension of the variance matrix of the coefficients is not limited to stochastic volatility VARs, but can happen also in a homoskedastic setting. In particular, consider making the model (1)-(2) homoskedastic:

$$y_t = \Pi_0 + \Pi(L)y_{t-1} + v_t, \tag{17}$$

$$v_t = A^{-1}\Lambda^{0.5}\epsilon_t,\ \epsilon_t \sim iid\ N(0, I), \tag{18}$$

triangular. Clearly, draws from $C^{-1} \times \text{rand}$ will have variance $\bar{\Omega}_\Pi$ so we can use $C^{-1} \times \text{rand}$ rather than $chol(\bar{\Omega}_\Pi) \times \text{rand}$. Moreover we can substitute $\bar{\Omega}_\Pi = C^{-1}C'^{-1}$ in (13) and take C^{-1} as common factor to obtain:

$$\text{vec}(\Pi^m) = C^{-1}\left[C^{-1\prime}\left\{\text{vec}\left(\sum_{t=1}^{T} X_t y_t' \Sigma_t^{-1}\right) + \underline{\Omega}_\Pi^{-1}\text{vec}(\underline{\mu}_\Pi)\right\} + \text{rand}\right]. \tag{15}$$

In the expression above, the computation of Π^m requires i) computing C', the Cholesky factor of $\bar{\Omega}_\Pi^{-1}$; ii) obtaining $C^{-1\prime}$ by inverting C'; iii) performing the two multiplications of the terms in the curly and square brackets by $C^{-1\prime}$ and C^{-1} respectively. However, in the above expression C is triangular so its inversion is less expensive, in particular one can simply use the command for backward solution of a linear system as suggested by Chan (2015) instead of inverting the matrices:

$$\text{vec}(\Pi^m) = C\backslash\left[C'\backslash\left\{\text{vec}\left(\sum_{t=1}^{T} X_t y_t' \Sigma_t^{-1}\right) + \underline{\Omega}_\Pi^{-1}\text{vec}(\underline{\mu}_\Pi)\right\} + \text{rand}\right], \tag{16}$$

where $X = C\backslash B$ is the matrix division of C into B, which is roughly the same as $C^{-1}B$, except it is computed as the solution of the equation $CX = B$. A draw in this case still requires the computation of the Cholesky factor of $\bar{\Omega}_\Pi^{-1}$ and its inversion, but the multiplications are avoided. Moreover in general computing inverse matrixes using the $\backslash$ operator is faster and more precise than matrix inversion in softwares such as Matlab. Therefore, using (16) to perform a draw requires only $2O(N^6)$. While this is twice as fast as using (13), it is just a linear improvement and it is not sufficient to solve the bottleneck in estimation of large systems, as the overall computational complexity for calculating a draw is still of the order $O(N^6)$. In the remainder of the paper we use the strategy outlined in this footnote for all the models we consider.

where the subscript t has been eliminated from the matrix Λ, so that we have

$$Var(v_t) = \Sigma = A^{-1}\Lambda A^{-1\prime}. \tag{19}$$

For this model, the prior distribution typically used is

$$\begin{aligned} \text{vec}(\Pi) &\sim N(\text{vec}(\underline{\mu}_\Pi), \underline{\Omega}_\Pi); & (20)\\ \Sigma &\sim IW(\underline{d}_\Sigma \cdot \underline{\Sigma}, \underline{d}_\Sigma), & (21) \end{aligned}$$

and the implied posteriors are

$$\begin{aligned} \text{vec}(\Pi)|\Sigma, y &\sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi); & (22)\\ \Sigma|\Pi, y &\sim IW((\underline{d}_\Sigma + T) \cdot \bar{\Sigma}, \underline{d}_\Sigma + T); & (23) \end{aligned}$$

with

$$\bar{\Omega}_\Pi^{-1} = \underline{\Omega}_\Pi^{-1} + \sum_{t=1}^{T} (\Sigma^{-1} \otimes X_t X_t'). \tag{24}$$

The matrix in (24) still has the same dimension of the one in (14), notwithstanding the fact that the matrix Σ does not vary with time.

The papers that have estimated homoskedastic VARs with a large cross section all use a different prior for Π:

$$\text{vec}(\Pi)|\Sigma \sim N(\text{vec}(\underline{\mu}_\Pi), \Sigma \otimes \Omega_0), \tag{25}$$

that is, the prior is conditional on knowledge of Σ, and the matrix Σ is used to elicit the prior variance $\underline{\Omega}_\Pi = \Sigma \otimes \Omega_0$. Under these assumptions equation (24) simplifies to:

$$\bar{\Omega}_\Pi^{-1} = \Sigma \otimes \left\{ \Omega_0 + \sum_{t=1}^{T} X_t X_t' \right\}, \tag{26}$$

which has a Kronecker structure that permits manipulating the two terms in the Kronecker product separately (for details see Carriero, Clark and Marcellino 2015), which provides huge computational gains and reduces the complexity to N^3. This specification allowed researchers, starting with Banbura, Giannone and Reichlin (2010), to estimate BVARs with more than a hundred variables.

However, a specification such as (25) is restrictive, as highlighted by Zellner (1973), Kadiyala and Karlsson (1997), because it prevents permitting any asymmetry in the prior across equations, and it requires specifying the prior on the mean coefficients conditionally on the prior on the variance coefficients. For example, the traditional Minnesota prior in the original Litterman (1986) implementation can not be cast in such a convenient form, because it imposes extra shrinkage on lags of variables that are not the lagged dependent

variable in each equation. Moreover, as noted by Sims and Zha (1998), the restriction $\underline{\Omega}_{\Pi} = \Sigma \otimes \Omega_0$, which is necessary to preserve a convenient system structure, implies the unappealing consequence that prior beliefs are correlated across equations of the reduced form representation of the VAR, with a correlation structure proportional to that of the disturbances.[4]

As we shall see, our proposed algorithm also solves the prior asymmetry problem and allows the estimation — for example — of a large VAR with the traditional Minnesota prior and random error variance.

2.4 The triangular algorithm

In this section we propose a very simple algorithm which solves the problems we discussed in the previous subsections. The algorithm does so simply by blocking the conditional posterior distribution in (10) in N different blocks. Recall that in the step of the Gibbs sampler that involves drawing Π, all of the remaining model coefficients are given, and consider again the decomposition $v_t = A^{-1}\Lambda_t^{0.5}\epsilon_t$:

$$\begin{bmatrix} v_{1,t} \\ v_{2,t} \\ ... \\ v_{N,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & ... & 0 \\ a^*_{2,1} & 1 & & ... \\ ... & & 1 & 0 \\ a^*_{N,1} & ... & a^*_{N,N-1} & 1 \end{bmatrix} \begin{bmatrix} h^{0.5}_{1,t} & 0 & ... & 0 \\ 0 & h^{0.5}_{2,t} & & ... \\ ... & & ... & 0 \\ 0 & ... & 0 & h^{0.5}_{N,t} \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ ... \\ \epsilon_{N,t} \end{bmatrix}, \tag{27}$$

where $a^*_{j,i}$ denotes the generic element of the matrix A^{-1} which is available under knowledge of A. The VAR can be written as:

$$\begin{aligned} y_{1,t} &= \pi_1^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{1,l}^{(i)} y_{i,t-l} + h_{1,t}^{0.5}\epsilon_{1,t} \\ y_{2,t} &= \pi_2^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{2,l}^{(i)} y_{i,t-l} + a^*_{2,1}h_{1,t}^{0.5}\epsilon_{1,t} + h_{2,t}^{0.5}\epsilon_{2,t} \\ &... \\ y_{N,t} &= \pi_N^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{N,l}^{(i)} y_{i,t-l} + a^*_{N,1}h_{1,t}^{0.5}\epsilon_{1,t} + \cdots + a^*_{N,N-1}h_{N-1,t}^{0.5}\epsilon_{N-1,t} + h_{N,t}^{0.5}\epsilon_{N,t}, \end{aligned}$$

[4] Sims and Zha (1998) propose an approach which allows for a more general structure of the coefficient prior variance, and which attains computational gains also of order $O(N^2)$. However, their approach is restricted to homoskedastic VARs and is based on the *structural* equations of the system. In particular, their prior achieves computational gains by assuming independence across the coefficients belonging to different structural equations, but the implied correlations across reduced form coefficients are still proportional to the correlations of the disturbances. For this reason their approach can not achieve computational gains for an asymmetric prior on the reduced form equations coefficients, as explained in section 5.2 of their paper.

with the generic equation for variable j:

$$y_{j,t} - (a^*_{j,1} h^{0.5}_{1,t} \epsilon_{1,t} + ... + a^*_{j,,j-1} h^{0.5}_{j-1,t} \epsilon_{j-1,t}) = \pi^{(0)}_j + \sum_{i=1}^{N} \sum_{l=1}^{p} \pi^{(i)}_{j,l} y_{i,t-l} + h_{j,t} \epsilon_{j,t}. \quad (28)$$

Consider estimating these equations in order from $j = 1$ to $j = N$. When estimating the generic equation j the term of the left hand side in (28) is known, since it is given by the difference between the dependent variable of that equation and the estimated residuals of all the previous $j - 1$ equations. Therefore we can define:

$$y^*_{j,t} = y_{j,t} - (a^*_{j,1} h^{0.5}_{1,t} \epsilon_{1,t} + ... + a^*_{j,,j-1} h^{0.5}_{j-1,t} \epsilon_{j-1,t}), \quad (29)$$

and equation (28) becomes a standard generalized linear regression model for the variable in equation (29) with i.i.d. Gaussian disturbances with mean 0 and variance $h_{j,t}$. The distribution (10) can be factorized as:

$$\begin{aligned} p(\Pi|A, \Phi, \Lambda_T, y) &= p(\pi_{N,1}, ..., \pi_{N,N}|\pi_{N-1,1}, ..., \pi_{N-1,N}, ..., \pi_{1,1}, ..., \pi_{1,N}, A, \Lambda_T, y) \\ &\quad \times p(\pi_{N-1,1}, ..., \pi_{N-1,N}|\pi_{N-2,1}, ..., \pi_{N-2,N}, ..., \pi_{1,1}, ..., \pi_{1,N}, A, \Lambda_T, y) \\ &\quad \times ... \times p(\pi_{1,1}, ..., \pi_{1,N}|A, \Phi, \Lambda_T, y). \end{aligned} \quad (30)$$

Using the factorization in (30) together with the model in (28) allows to draw the coefficients of the matrix Π in separate blocks. Define the j-th row of the matrix Π as $\Pi^{\{j\}} = \{\pi_{j1}, ..., \pi_{jN}\}$ and all the previous rows as $\Pi^{\{1:j-1\}}$. Then draws of $\Pi^{\{j\}}$ can be obtained from:

$$\Pi^{\{j\}}|\Pi^{\{1:j-1\}}, A, \Lambda_T, y \sim N(\bar{\mu}_{\Pi^{\{j\}}}, \overline{\Omega}_{\Pi^{\{j\}}}) \quad (31)$$

with

$$\bar{\mu}_{\Pi^{\{j\}}} = \overline{\Omega}_{\Pi^{\{j\}}} \left\{ \sum_{t=1}^{T} X_{j,t} h^{-1}_{j,t} y^{*\prime}_{j,t} + \underline{\Omega}^{-1}_{\Pi^{\{j\}}} (\underline{\mu}_{\Pi^{\{j\}}}) \right\} \quad (32)$$

$$\overline{\Omega}^{-1}_{\Pi^{\{j\}}} = \underline{\Omega}^{-1}_{\Pi^{\{j\}}} + \sum_{t=1}^{T} X_{j,t} h^{-1}_{j,t} X'_{j,t} \quad (33)$$

where $y^*_{j,t}$ is defined in (29) and where $\underline{\Omega}^{-1}_{\Pi^{\{j\}}}$ and $\underline{\mu}_{\Pi^{\{j\}}}$ denote the prior moments on the j-th equation, given by the j-th column of $\underline{\mu}_{\Pi}$ and the j-th block on the diagonal of $\overline{\Omega}^{-1}_{\Pi}$. Note we have implicitly assumed here that the matrix $\underline{\Omega}^{-1}_{\Pi}$ is block diagonal, which means that we are ruling out any prior correlation among the coefficients belonging to different equations. This is a restriction with respect to the more general model, however we note

that the typical priors elicited in the literature for the matrix Ω typically do not involve cross-equation correlations.[5]

The dimension of the matrix $\overline{\Omega}_{\Pi\{j\}}^{-1}$ is $(Np+1)$, which means that its manipulation only involves operations of order $O(N^3)$. However, since in order to obtain a draw for the full matrix Π one needs to draw separately all of its N rows, the total computational complexity of this algorithm is $O(N^4)$. This is considerably smaller than the complexity of $O(N^6)$ implied by the standard algorithm, with a gain of N^2. For a model with 20 variables this difference amounts to a 400-fold improvement in estimation time. Where is the computational gain coming from? In the traditional algorithm the sparsity implied by the possibility of triangularising the system is not exploited, and all computations are carried out using the whole vectorized system. In this algorithm instead the triangularization allows to estimate equations which are at most containing $Np+1$ regressors, and the correlation among the different equations typical of SUR models is implicitly accounted for by the triangularisation scheme.

Finally, note that in a homoskedastic model the same reasoning for drawing the coefficients Π applies, so that the relevant posterior distributions for the Gibbs sampler would again be given by equation (31), with prior mean and variance given by formulas (32) and (33), with the only difference being that the subscript t would be omitted from the volatility terms $h_{j,t}$. For this reason, the equation-by-equation step can be also used to estimate large VARs with asymmetric priors, such as, e.g., the Minnesota prior.

In closing this Subsection it is worth to stress that expression (27) and the following triangular system are based on a Cholesky-type decomposition of the variance Σ_t, but such decomposition here is simply used as an estimation device, not as a way to identify structural shocks. The ordering of the variables in the system does not change the joint posterior of the reduced form coefficients, so changing the order of the variables is inconsequential to the results, even though it is of course convenient to order the variables in a way that is already consistent with the preferred strategy for identification of structural shocks.

2.5 MCMC samplers

To conclude, we summarize the steps involved in the MCMC samplers for the BVAR with stochastic volatility and for a BVAR with asymmetric priors, highlighting how all the existing

[5] A notable exception is the conjugate prior for a homoskedastic VAR in (25), which restricts the structure of the correlations to be proportional to the error variance. Some priors involve prior correlations among coefficients of the same equations, notably the sum of coefficients and unit root prior proposed by Sims (1993) and Sims and Zha (1998). This case is still consistent with our block-diagonal specification for the matrix $\underline{\Omega}_{\Pi}^{-1}$.

algorithms can be easily modified to include our equation-by-equation step in place of the standard system-wide step for drawing the VAR conditional mean coefficients.

2.5.1 Gibbs sampler for large VAR with stochastic volatility

We estimate the BVAR model with stochastic volatility (BVAR-SV) with a Gibbs sampler. Let s^T denote the states of the mixture of normals distribution used in the Kim, Shephard, and Chib (1998) algorithm, and recall that Θ denotes all the model coefficients, while y_T and Λ_T denote the full time series of the data and states.

The Gibbs sampler draws in turn from the conditionals $p(\Lambda_T|\Theta, s^T, y_T)$ and $p(\Theta, s^T|\Lambda_T, y_T)$.

Step 1: Draw from $p(\Lambda_T\ |\Theta, s^T, y_T)$ relying on the state space representation described above and the Kalman filter and simulation smoother of Durbin and Koopman (2001).

Step 2: Draw from $p(\Theta, s^T|\Lambda_T, y_T)$ relying on the factorization $p(\Theta, s^T|\Lambda_T, y) \propto p(s^T|\Theta, \Lambda_T, y) \cdot p(\Theta|\Lambda_T, y)$, that is by (i) drawing from the marginal posterior of the model parameters $p(\Theta|\Lambda_T, y_T)$ and (ii) drawing from the conditional posterior of the mixture states $p(s^T|\Theta, \Lambda_T, y_T)$. The marginal posterior $p(\Theta|\Lambda_T, y_T)$ is sampled by further breaking the parameter block into pieces and drawing from the distributions of each parameter piece conditional on the other parameter pieces (steps 2a-2c below), while draws from $p(s^T|\Theta, \Lambda_T, y_T)$ (step 2d) are obtained using steps similar to those described in Primiceri (2005). In more detail, the sub-steps used to produce draws from $p(\Theta, s^T|\Lambda_T, y_T)$ are as follows.

Step 2a: Draw Φ conditional on the data and Λ_T, using the conditional (IW) distribution for the posterior given in (12).

Step 2b: Draw the matrix of VAR coefficients Π *equation by equation*, conditional on the data, A and Λ_T, using the conditional (normal) distribution for the posteriors given in equation (31) and the factorization (30).

Step 2c: Draw the elements of the matrix A conditional on the data, Π and Λ_T, using the conditional distribution for the posterior given in (11).

Step 2d: Draw the states of the mixture of normals distribution s^T conditional on the data, Λ_T, and the parameter block Θ.

Alternatively, if the innovations to volatility are assumed to be uncorrelated, one can use the Cogley and Sargent (2005) approach to draw the volatility states Λ_T. In such case there is no need to introduce the mixture states s^T and therefore step 2d is not necessary while step 1 uses an independence Metropolis step such as the one described in Cogley and Sargent (2005).

Note that the only difference between this algorithm and the standard algorithm used in most implementations of VARs with stochastic volatility is in step 2b, which here is

performed equation by equation. This means that if a researcher already has a standard algorithm, its computational efficiency can be easily improved by simply replacing the traditional system wide step to draw Π with step 2b.

2.5.2 Gibbs sampler for large VAR with asymmetric prior

In the case of a homoskedastic model with an asymmetric prior the Gibbs sampler works as follows.

Step 1: Draw the matrix of VAR coefficients Π *equation by equation*, conditional on the data, A, and Λ using the conditional (normal) distribution given in equation (31) and the factorization (30).

Step 2: Draw the matrix Σ conditional on the data and Π, using the conditional (IW) distribution for the posterior given in (23), and derive the matrices A^{-1} and Λ using the decomposition in equation (19).

Note that the only difference between this algorithm and the standard algorithm used e.g. in Kadiyala and Karlsson (1997) for the independent Normal-Wishart prior is in step 1, which here is performed equation by equation. This means that if a researcher already has a standard algorithm, its computational efficiency can be easily improved by simply replacing the traditional system-wide step to draw Π with step 1 above.

3 A numerical comparison of the estimation methods

In this section we compare the proposed triangular algorithm with the traditional system-wide algorithm for estimation of the VAR in (1)-(2).

3.1 Computational complexity and speed of simulation

First, we compare the results obtained by using either algorithm as the dimension of the cross section N increases. We use data taken from the dataset of McCracken and Ng (2015) (MN dataset), at monthly frequency, from January 1960 to December 2014. The data are transformed as in McCracken and Ng (2015) to achieve stationarity and their short acronyms are listed in Table 1.

We start by simply comparing the posterior estimates obtained using the two alternative algorithms, focussing on a medium-sized system of 20 variables and 13 lags. The 20 variables we select for this exercise are identified by a star in Table 1, and they include a selection of the most relevant time series in the MN dataset. Figure 1 presents the impulse response functions to a monetary policy shock defined as a shock to the federal funds rate obtained

using the two alternative algorithms, based on 5000 draws from the posterior distribution after 500 draws of burn-in. Of course, the two algorithms produce the same results, and any residual difference is due to sample variation and is bound to disappear as the number of replication increases.[6] A similar picture comparing the (time series of) the distributions of the time-varying volatilities shows completely indistinguishable results, and for this reason we omit it.

Importantly, though, the estimation of the model using the traditional system-wide algorithm was about 261 times slower. This represents a substantial improvement in the ease of estimating and handling these models, which is relevant especially in consideration of the fact that models of this size have been markedly supported by the empirical evidence in contributions such as Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015) and Koop (2013). Figure 2 illustrates the computational gains arising from the use of the triangular algorithm. The top panel shows the computational time (on a 2014 top-range iMac) needed to perform 10 draws as a function of the size of the cross section using the triangular algorithm (green line) and the system-wide algorithm (blue line). As is clear, the computational gains grow nonlinearly and become already substantial with $N > 5$. The bottom panel compares the gain in theoretical computational complexity (black dashed line - which is equal to N^2) with the actual computational time. As is clear, for smaller systems the computational gains achieved are below the theoretical ones, but this is due to all the other operations involved in the estimation rather than the core computations involving the inversion of the coefficients posterior variance matrix.

In order to explore what happens for cross sections larger than $N = 10$, Figure 3 extends the results of Figure 2 up to $N = 40$. These results are computed by including additional variables from the MN dataset. Since the computational gains become so large that they create scaling problems, results in this Figure are displayed using a logarithmic vertical axis. As is clear, the computational gains from the triangular algorithm grow quadratically, and after $N = 25$ they become even larger than the theoretical gains, which we attribute to the fact that for such large systems the size of the operations is so large that it saturates the CPU computing power.

[6] We repeated the exercise shutting down the random variation, i.e. using exactly the same random seed for the two algorithms, and the results exactly coincide besides minimal numerical errors.

3.2 Convergence and mixing

Clearly, as shown in Figure 1, the traditional step-wise and the proposed triangular algorithm produce draws from the same posterior distribution. It could be argued that - as long as we have an increasing computing power - using the triangular algorithm only achieves gains in terms of speed. However, it is important to stress that - regardless of the power or the computers used to perform the simulation - the triangular algorithm will always produce many more draws than the traditional system-wide algorithm in the same unit of time. This has important consequences in terms of producing draws with good mixing and convergence properties.

To illustrate this point, we consider the quality of the draws that we can obtain from the two algorithms *within a given amount of time*. Specifically, for the 20-variable model with Minnesota prior and stochastic volatility described in the previous subsection, we first run the system-wide algorithm and produce 5000 draws from it and record the total time needed to produce these draws. Then, we run the triangular algorithm for the same amount of time, and out of all the draws produced in this time interval, which are 261 times more -since this algorithm is about 261 times faster, we perform skip-sampling by saving only each 261-th draw. Obviously, this results in the same number of final draws (5000) but these draws have dramatically improved convergence and mixing properties. Figure 4 plots the Inefficiency Factors of 5000 draws obtained by running the two alternative algorithms *for the same amount of time*. As is clear, the Inefficiency Factors produced by the triangular algorithm are way lower than those obtained by the system-wide algorithm. The triangular algorithm can produce draws many times closer to i.i.d. sampling in the same amount of time. Being closer to i.i.d sampling, the draws from the triangular algorithm feature better convergence properties. Instead, the system-wide algorithm is slower to converge (in a unit of time), especially so for the coefficients related to volatility (the innovations to volatility and the volatility states). Figure 5 illustrates the recursive means for some selected coefficients and shows that the triangular algorithm with split sampling reaches convergence much faster than the system-wide algorithm, and this pattern is particularly marked for the volatility component of the model.

Since these gains are increasing nonlinearly with the system size, we conclude that, for conventional forecasting or structural analysis with medium and large BVARs, the triangular algorithm offers computational gains large enough that many researchers should find it preferable. This should be especially true in forecasting analyses that involve model estimation at many different points in time.

4 Example: A large VAR with drifting volatilities

In this Section we provide an example of how the triangular algorithm can be used to estimate a very large BVAR with drifting volatilities. We consider a VAR with 125 variables, which includes all of the variables considered by McCracken and Ng (2015) with the exception of housing permits and their disaggregate components, which we exclude since these variables produced problems of collinearity.

We use a specification with 13 lags and the prior mean and variance of the coefficients set using an independent Normal-Wishart prior which reflects the prior mean and variances of the original Minnesota prior. This means that we do impose cross-variable shrinkage. Finally, all of the errors feature stochastic volatility. The total number of objects to estimate is given by 203250 mean coefficients, 7750 covariance coefficients, and 125 latent states (each of length T). Despite the huge dimension of the system, the proposed algorithm can produce 5000 draws (after 500 of burning in) in just above 7 hours on a 2014 top-of the range iMac.

Results are summarized in Figures 6-10. Figure 6 provides convergence diagnostics (Inefficiency Factors and Potential Scale Reduction Factors) on the various parameters and latent states. As is clear from the figure, once a skip-sampling of 5 is performed (leaving 1000 clean draws) the algorithm has good convergence and mixing properties. Note that, with a model this large, skip-sampling greatly reduces storage costs.

Figures 7 and 8 present the estimated volatilities. It turns out that there is substantial homogeneity in the estimated volatility patterns for variables belonging to the same group, such as IP and PPI components or interest rates at different maturities, but there is some heterogeneity across groups of variables. Moreover, while the Great Moderation starting around 1985 is evident in most series, the effects of the recent crisis are more heterogeneous. In particular, while volatility of real variables, such as IP and employment, and financial variables, such as stock price indexes, interest rates and spreads, goes back to lower levels after the peak associated with the crisis, there seems to remain a much higher level of volatility than before the crisis in price indicators, in particular in PPI and its components and also in several CPI components as well as in monetary aggregates, but also in housing starts. Overall, the first principal component of all the estimated volatilities explains about 56% of overall variance, and the first three 85%, confirming that commonality is indeed present but idiosyncratic movements also matter (as in the GFSV specification of Carriero et al. (2012)).

Figures 9 and 10 present the estimated impulse response functions to a unitary shock to the federal funds rate, replicating in our context the analysis of Bernanke, Boivin and Eliasz (2005), based on a constant parameter FAVAR, and that of Banbura, Giannone and

Reichlin (2010) based on a large VAR with homoskedastic errors. For identification, the federal funds rate is ordered after slow-moving and before fast-moving variables. A first, and obvious, comment is that the size of the shock was clearly not stable over time, as from Figure 8 the volatility of the federal funds rate changed substantially over time, so that the overall contribution of the monetary policy shock is also changing over time, while it is assumed constant in models with homoskedastic errors. Next, looking at all the responses, we see that they look reasonable, with a significant deterioration in real variables such as IP, unemployment, employment and housing starts, only very limited evidence of a price puzzle, with most price responses not statistically significant, a significant deterioration in stock prices, a less than proportional increase in the entire term structure, which leads to a decrease in the term spreads, progressively diminishing over time, and a negative impact on the ISM indexes. Overall, the responses are in line with those reported in Banbura, Giannone and Reichlin (2010) since, as we have seen, the presence of heteroskedasticity does not affect substantially the VAR coefficient estimates, but it matters for calculating the confidence bands and understanding the evolution of the size of the shock (and therefore of the responses) over time. Stochastic volatility would also matter for variance decompositions, omitted here in the interest of brevity.

5 Conclusions

In this paper we have proposed a new algorithm to perform estimation of large VARs with possibly asymmetric priors and drifting volatilities. The algorithm is based on a straightforward triangularization of the system, and it is very simple to implement. The algorithm ensures computational gains of order N^2 with respect to the traditional algorithm used to estimate VARs with independent Normal-Wishart priors, and because of this it is possible to achieve much better mixing and convergence properties compared to existing algorithms. We have illustrated the algorithm with an empirical application on the effects of a monetary policy shock in a large Vector Autoregression. Given its simplicity and the advantages in terms of speed, mixing, and convergence, we argue that the proposed algorithm should be preferred in empirical applications involving large datasets.

References

[1] Banbura, M., Giannone, D., and Reichlin, L., 2010. Large Bayesian Vector Autoregressions, *Journal of Applied Econometrics* 25, 71-92

[2] Bernanke, B., J. Boivin, P. Eliasz, 2005, 'Measuring the effects of monetary policy: a Factor-Augmented Vector Autoregressive (FAVAR) approach', The Quarterly Journal of Economics, 120(1), 387-422.

[3] Carriero A., Clark, T. and Marcellino, M., 2012. Common Drifting Volatility in Large Bayesian VARs. *Journal of Business and Economic Statistics*, forthcoming.

[4] Carriero A., Clark, T. and Marcellino, M., 2015. Bayesian VARs: Specification Choices and Forecast Accuracy. *Journal of Applied Econometrics*, 30, 46-73.

[5] Chan, J., 2015. Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure, manuscript, 2015.

[6] Clark, T., 2011. Real-Time Density Forecasts from BVARs with Stochastic Volatility, *Journal of Business and Economic Statistics* 29, 327-341.

[7] Clark, T., and Ravazzolo, F., 2015. Macroeconomic Forecasting Performance Under Alternative Specifications of Time-Varying Volatility, *Journal of Applied Econometrics*, 30, 551-575.

[8] Cogley, T., and Sargent, T., 2005. Drifts and Volatilities: Monetary Policies and Outcomes in the post-WWII US, *Review of Economic Dynamics* 8, 262-302.

[9] D'Agostino, D., Gambetti, L., and Giannone, D., 2013. Macroeconomic forecasting and structural change, *Journal of Applied Econometrics* 28, 82-101.

[10] Del Negro, M., and Primiceri, G., 2014. Time-Varying Structural Vector Autoregressions and Monetary Policy: A Corrigendum, manuscript, Northwestern University.

[11] Durbin, J. and Koopman, S.J., 2001, *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford, UK.

[12] Giannone, D., Lenza, M. and G. Primiceri, 2015. Prior Selection for Vector Autoregressions. *Review of Economics and Statistics*, 97(2), 436-451.

[13] Jacquier, E., Polson, N.G., Rossi, P. E., 2002, Bayesian Analysis of Stochastic Volatility Models. *Journal of Business & Economic Statistics* 20(1), 69-87.

[14] Kadiyala, K., and Karlsson, S., 1997. Numerical Methods for Estimation and Inference in Bayesian VAR-Models, *Journal of Applied Econometrics* 12, 99-132.

[15] Kim, S., Shephard, N. and S. Chib, 1998. Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies* 65, 361-393.

[16] Koop, G., 2013. Forecasting with Medium and Large Bayesian VARs, *Journal of Applied Econometrics* 28, 177-203.

[17] Koop, G., and Korobilis, D., 2013. Large Time-Varying Parameter VARs. *Journal of Econometrics* 177, 185-198.

[18] Litterman, R., 1986. Forecasting with Bayesian Vector Autoregressions-Five Years of Experience, *Journal of Business and Economic Statistics* 4, 25-38.

[19] McCracken, M.W., Ng, S., 2015. FRED-MD: A Monthly Database for Macroeconomic Research. Working Papers 2015-12, Federal Reserve Bank of St. Louis.

[20] Primiceri, G., 2005. Time Varying Structural Vector Autoregressions and Monetary Policy, *Review of Economic Studies* 72, 821-852.

[21] Sims, C., 1993. A Nine-Variable Probabilistic Macroeconomic Forecasting Model, in *Business Cycles, Indicators and Forecasting*, James H. Stock and Mark W. Watson, editors, University of Chicago Press, 179-212.

[22] Sims, C., and Zha, T., 1998. Bayesian Methods for Dynamic Multivariate Models, *International Economic Review* 39, 949-68.

[23] Zellner A. 1973. An Introduction to Bayesian Inference in Econometrics. Wiley: New York.