Large Vector Autoregressions with stochastic volatility and non-conjugate priors*

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: June 2017

Abstract

Recent research has shown that a reliable Vector Autoregressive model (VAR) for forecasting and structural analysis of macroeconomic data requires a large set of variables and modeling time variation in their volatilities. Yet, there are no papers jointly allowing for stochastic volatilities and large datasets, due to computational complexity. Moreover, homoskedastic VAR models for large datasets so far restrict substantially the allowed prior distributions on the parameters. In this paper we propose a new Bayesian estimation procedure for (possibly very large) VARs featuring time varying volatilities and general priors. This is important both for reduced form applications, such as forecasting, and for more structural applications, such as computing response functions to structural shocks. We show that indeed empirically the new estimation procedure performs very well for both tasks.

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

*We would like to thank Joshua Chan, Ana Galvao, Gary Koop, Dimitris Korobilis, Haroon Mumtaz, Davide Pettenuzzo, Anna Simoni and participants at seminars and conferences at the Banque de France, ECB, Bank of England, and the University of Pennsylvania conference on Big Data in Predictive Dynamic Econometric Modeling for useful comments on a previous version. 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. The only exceptions are Koop and Korobilis (2013), Koop, et al. (2016), and Carriero, Clark, and Marcellino (2016). Koop and Korobilis (2013) and Koop, et al. (2016) propose a computational (not fully Bayesian) shortcut that allows for time-varying volatility using, roughly speaking, 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, computing the uncertainty around the volatility estimates in a coherent fashion. Our previous work in Carriero, Clark, and Marcellino (2016) 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 (2016) 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 cannot easily be estimated in a large VAR — without restrictions such as those of Carriero, Clark, and Marcellino (2016) — 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 as the number of variables increases. Homoskedastic

VAR models are SUR models 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 vectorized 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 where, even in the 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.[1] 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 cannot 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 the presence of a symmetric likelihood, the resulting posterior is not symmetric across equations, which implies that the system needs to be vectorized 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 one to estimate his model equation by equation.

To summarize, if either the prior or the likelihood induces an asymmetry in the posterior of the VAR coefficients, the model needs to be vectorized and its computational complexity rises from N^3 up to N^6, where N is the size of the cross section. This applies, for example,

[1] Studies including Chib and Greenberg (1995) and Korobilis and Pettenuzzo (2017) have developed approaches for large VARs (without stochastic volatility) that rely on hierarchical priors.

with the common Normal-diffuse and independent Normal-Wishart priors. For this reason the only VAR which can be reasonably estimated with a large cross section of data is the homoskedastic VAR with natural conjugate prior proposed by Kadiyala and Karlsson (1993, 1997),[2] which features symmetry in both the prior and the likelihood, and it is indeed this model upon which papers such as Banbura, Giannone, and Reichlin (2010) and Carriero, Clark, and Marcellino (2016) are built.

In this paper we propose a new estimation procedure that allows one to estimate VARs featuring asymmetries either in the prior or in the likelihood, thereby allowing for models with asymmetric priors and time varying volatilities. Our procedure is based on a simple triangularization of the VAR, which allows one to simulate the VAR coefficients by drawing them equation by equation. This 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 algorithms, and therefore it allows one to estimate large models. Moreover, our new algorithm is very simple and, importantly, it can be easily inserted in any pre-existing algorithm for estimation of BVAR models.

With our method, estimation of very large VARs with stochastic volatility becomes feasible, and this is important both for reduced form applications, such as forecasting or constructing coincident and leading indicators, and for more structural applications, such as computing response functions to structural shocks or forecast error variance decompositions. Hence, our method also paves the way for a large number of empirical applications.

As an example and illustration, we estimate a VAR with stochastic volatilities, using a cross-section of 125 variables for the U.S. extracted from the dataset in McCracken and Ng (2016).

A first interesting finding is 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

[2]The conjugate Normal-Inverse Wishart prior is discussed in Rothenberg (1963), and Zellner (1973) in the general context of multivariate regressions. Kadiyala and Karlsson (1993, 1997) proposed and studied this prior in the specific context of VARs. Geweke and Whiteman (2006) and Karlsson (2013) offer excellent surveys on priors for VARs.

well as in monetary aggregates and housing starts. Overall, the first principal component of all the estimated volatilities explains about 45% of overall variance, and the first three 73%, confirming that commonality is indeed present but idiosyncratic movements also matter (as in the GFSV specification of Carriero, Clark, and Marcellino (2016) and in the factor volatility model of Carriero, Clark and Marcellino (2017)).

Next, we use this very large VAR-SV to analyze US monetary policy shocks and their transmission, replicating the analysis of Bernanke, Boivin and Eliasz (2005, BBE), based on a constant parameter FAVAR, and that of Banbura, Giannone and Reichlin (2010, BGR), based on a large VAR with homoskedastic errors. Besides the common advantages of using large datasets in VARs, such as reducing the likelihood of omitted variables and non-fundamental shocks, we can now also allow for time-varying variances of the structural shocks. Indeed, a first result, perhaps obvious but omitted in previous analyses with large datasets such as BBE and BGR, is that the variance of the shocks was clearly unstable over time, so that the overall explanatory contribution of the monetary policy shocks is also changing over time, while it is assumed constant in models with homoskedastic errors. Next, we get a granular view of the dynamic propagation of the monetary shock: most of the 125 responses 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.

Finally, we analyze the effect that the size of the cross-section and the time variation in the volatilities has on out-of-sample forecasting performance. We compare small and medium sized (20 variable) VARs for the US, with and without stochastic volatility, in a recursive out-of-sample exercise, where the inclusion of the medium sized VAR-SV is only feasible thanks to our new estimation method. A priori, we expect the inclusion of time variation in volatilities to improve density forecasts, via a better modeling of error variances, while the use of a larger dataset should improve point forecasts, via a better specification of the conditional means. However, this is not the whole story, as there are also interaction effects: a better point forecast should improve density forecasts as well, by centering the predictive density around a more reliable mean, and time varying volatilities should improve the point forecasts — especially at longer horizons — because the heteroskedastic model will provide more efficient estimates (through a GLS argument) and a therefore a better characterization of the predictive densities, with the predictive means gradually deviating

from their homoskedastic counterparts as the predictive densities cumulate nonlinearly with the forecast horizon. Indeed this is precisely the pattern we find in the data, which confirms the usefulness of large VAR-SVs in a forecasting context.

The paper is structured as follows. Sections 2 and 3 introduce the model and develop the estimation method. Section 4 presents a numerical comparison to illustrate the gains in terms of computing time (and convergence and mixing properties). Section 5 discusses the identification of the monetary policy shock and its propagation in the very large VAR with time varying volatilities. Section 6 presents the out-of-sample forecasting exercise. Section 7 summarizes the main findings and concludes.

2 Challenges in estimating large VARs with asymmetric priors and time varying volatilities

2.1 The model

Consider the following VAR model with stochastic volatility:

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

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

where $t = 1, ..., T$, the dimension of the vectors y_t, v_t and ϵ_t is N, $\Pi(L) = \Pi_1 L + \Pi_2 L^2 + ... + \Pi_p L^p$, Λ_t is a diagonal matrix with generic j-th element $\lambda_{j,t}$ 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:

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

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 $\tilde{v}_{j,t} = \lambda_{j,t}^{0.5}\epsilon_{jt}$. Taking logs of squares of $\tilde{v}_{j,t}$ yields the following set of observation equations:

$$\ln \tilde{v}_{j,t}^2 = \ln \lambda_{j,t} + \ln \epsilon_{j,t}^2,\ j = 1, \ldots, N. \quad (4)$$

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

$$\ln \lambda_{j,t} = \ln \lambda_{j,t-1} + e_{j,t},\ j = 1, \ldots, N, \quad (5)$$

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 as in Primiceri (2005) and not diagonal as in Cogley

and Sargent (2005).[3]

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 as we did in robustness checks of Carriero, et al. (2017) with larger models), 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 A^{-1} as time varying, this would not impact the main computational advantage arising from the estimation method we 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. That said, although our proposed approach solves the main bottleneck due to the size of the variance matrix of the VAR coefficients, in large systems (e.g., 30 or more variables), the estimation of a time varying A matrix would still be challenging. For example, with a 30 variable model, the number of states present in the A matrix would be 435.[4] Finally, as a simpler or less computationally challenging matter, one can modify equation (5) so that the states $\ln \lambda_{j,t}$ follow an autoregressive process rather than a random walk, but again this is not essential to the main 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. The matrix Π collects the lag matrices $\Pi_0, \Pi_1, ..., \Pi_p$. Under the conventional system approach, the independent priors for the coefficient blocks of the model are as follows:

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

[3]The specification of Primiceri (2005) is more general and allows for the volatilities to be hit by a common shock (while their conditional means are modeled independently one another). However, as N gets large with respect to T, allowing correlations across variables might become problematic. In the case of a full Φ matrix, innovations to the volatility are modeled with an inverse Wishart prior, which needs to use at least $N+2$ degrees of freedom to be proper. With large N, this makes the prior highly informative, more so with quarterly data than monthly. If some researcher were worried about that, he/she could treat the innovations as independent and draw them from individual inverse gamma distributions, as in Cogley and Sargent (2005). Of course this amounts to imposing the restriction that both the prior and the likelihood have a diagonal Φ matrix, which can be seen as an even more informative prior than the Wishart one.

[4]There is also a problem related on how to calibrate the prior on so many state variables under an approach like that of Primiceri (2005), since pre-sample data are very limited.

The model is completed by eliciting a prior for the initial value of the state variables Λ_t, which we set to an uninformative Gaussian.

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:

$$\text{vec}(\Pi)|A, \Lambda_T, y_T \sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi); \quad (10)$$

$$A|\Pi, \Lambda_T, y_T \sim N(\bar{\mu}_A, \overline{\Omega}_A); \quad (11)$$

$$\Phi|\Lambda_T, y_T \sim IW((\underline{d}_\Phi + T) \cdot \bar{\Phi}, \underline{d}_\Phi + T), \quad (12)$$

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.[5]

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 a mixture of normals approximation and multi-move algorithm proposed by Kim, Shepard and Chib (1998).[6] 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 vectorized 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

[5]Note that knowledge of 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.

[6]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 (3.2.1) below. In the case of volatilities independent across equations one could instead use the single-move sampler of Jacquier, Polson and Rossi (1994) and avoid drawing the mixture states s_T.

standard Gaussian), denoted rand, and to compute:

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

where $X_t = [1, y_{t-1}', ..., y_{t-p}']'$ is the $(Np+1)$-dimensional vector collecting the regressors in equation (1). 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.[7]

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 (2016), Giannone,

[7]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 triangular. Clearly, draws from $C^{-1} \times$ rand will have variance $\bar{\Omega}_\Pi$ so we can use $C^{-1} \times$ rand rather than $chol(\bar{\Omega}_\Pi) \times$ 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.

Lenza, and Primiceri (2015) and Koop (2013), this amounts to $4 \times 20^6 = 256$ million elementary operations (per single draw), and this is the main bottleneck that prevented the existing literature from estimating models with stochastic volatility using more than a handful of variables, typically 3 to 5.[8]

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 related to 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:

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

$$v_t \sim iid\ N(0, \Sigma). \tag{18}$$

Our VAR with stochastic volatility simplifies to this model if time variation in Λ_t is eliminated, as the Choleski decomposition of the constant Σ is equivalent to $A^{-1}\Lambda^{0.5}$.

For this model, the prior distribution typically used takes a Normal-diffuse or independent Normal-Wishart form (e.g., Karlsson 2013). Although our results also apply in the Normal-diffuse case, we will focus on the independent Normal-Wishart prior:

$$\text{vec}(\Pi) \sim N(\text{vec}(\underline{\mu}_\Pi), \underline{\Omega}_\Pi); \tag{19}$$

$$\Sigma \sim IW(\underline{d}_\Sigma \cdot \underline{\Sigma}, \underline{d}_\Sigma), \tag{20}$$

[8] Our previous work in Carriero, Clark, and Marcellino (2016) does estimate a larger system, but it does so by assuming a specific structure for the volatilities in the VAR. In particular, the matrix Σ_t in (14) is assumed to be given by the product of a scalar σ_t and a constant matrix Σ ($\Sigma_t = \sigma_t\Sigma$), and the prior variance $\underline{\Omega}_\Pi$ is specified conditionally on the error variance, $\underline{\Omega}_\Pi = \Sigma \otimes \Omega_0$, where the Kronecker product constrains the prior to be symmetric across equations. Under these restrictions, equation (14) can be written as $\bar{\Omega}_\Pi^{-1} = \Sigma^{-1} \otimes \{\Omega_0^{-1} + \sum_{t=1}^T \sigma_t^{-1} X_t X_t'\}$, which does have a Kronecker structure and therefore can be easily handled. However, the assumption $\Sigma_t = \sigma_t\Sigma$ imposes a specific factor structure on the volatilities which implies that all the conditional volatilities are driven by a single factor (σ_t) with a loading of 1, and there is no idiosyncratic component. This setup implies that the order of magnitude of the movements in volatility is proportional across variables. Although the evidence in Carriero, Clark, and Marcellino (2016) 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.

and the implied posteriors are

$$\begin{aligned} \text{vec}(\Pi)|\Sigma, y &\sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi); \quad (21) \\ \Sigma|\Pi, y &\sim IW((\underline{d}_\Sigma + T) \cdot \bar{\Sigma}, \underline{d}_\Sigma + T); \quad (22) \end{aligned}$$

with

$$\bar{\Omega}_\Pi^{-1} = \underline{\Omega}_\Pi^{-1} + \sum_{t=1}^{T} (\Sigma^{-1} \otimes X_t X_t'). \quad (23)$$

The matrix in (23) still has the same dimension as 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, of the conjugate Normal-Wishart form:

$$\text{vec}(\Pi)|\Sigma \sim N(\text{vec}(\underline{\mu}_\Pi), \Sigma \otimes \Omega_0). \quad (24)$$

In this case, 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 (23) simplifies to:

$$\bar{\Omega}_\Pi^{-1} = \Sigma^{-1} \otimes \left\{ \Omega_0^{-1} + \sum_{t=1}^{T} X_t X_t' \right\}, \quad (25)$$

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 (24) is restrictive, as highlighted by Rothenberg (1963), Zellner (1973), Kadiyala and Karlsson (1993, 1997), and Sims and Zha (1998), and there are many situations in which the form (24) can turn out to be particularly unappealing.

First, it prevents permitting any asymmetry in the prior across equations, because the coefficients of each equation feature the same prior variance matrix Ω_0 (up to a scale factor given by the elements of Σ). For example, the traditional Minnesota prior in the original Litterman (1986) implementation cannot 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. As another example, consider the case of a bivariate VAR in the variables y_1 and y_2 and suppose that the researcher has a strong prior belief that y_2 does not Granger cause y_1, while he/she has not strong beliefs that y_2 itself follows a univariate stationary process. This system of beliefs would require shrinking strongly towards zero the coefficients

attached to y_2 in the equation for y_1. However, in order to keep the conjugate structure (24) this would also necessarily require shrinking strongly towards their prior means also the coefficients attached to y_2 in the equation for y_2, and this is unpleasant since the researcher does not have such strong priors in this respect.

Second, the Kronecker structure $\Sigma \otimes \Omega_0$ in (24) also implies the unappealing consequence that prior beliefs must be correlated across the equations of the reduced form representation of the VAR, with a correlation structure proportional to that of the disturbances (as described by the matrix Σ). Sims and Zha (1998) discuss this issue in depth, and propose an approach which allows for a more reasonable structure of the coefficient prior variance, and which also attains — like our proposal below — computational gains of order $O(N^2)$. Their approach is based on eliciting a prior featuring independence among the *structural* equations of the system, but does not achieve computational gains for an asymmetric prior on the *reduced form* equations coefficients.[9]

As we shall see, our estimation method solves the problems outlined above, making the independent N-IW prior applicable in general, regardless of the size of the cross-section.

3 An estimation method for large VARs

In this section we propose a very simple estimation method that solves the problems we discussed above. It 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 $v_t = A^{-1}\Lambda_t^{0.5}\epsilon_t$:

$$\begin{bmatrix} v_{1,t} \\ v_{2,t} \\ ... \\ v_{N,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 & ... & 0 \\ a^*_{2,1} & 1 & & ... \\ ... & & 1 & 0 \\ a^*_{N,1} & ... & a^*_{N,N-1} & 1 \end{bmatrix} \begin{bmatrix} \lambda_{1,t}^{0.5} & 0 & ... & 0 \\ 0 & \lambda_{2,t}^{0.5} & & ... \\ ... & & ... & 0 \\ 0 & ... & 0 & \lambda_{N,t}^{0.5} \end{bmatrix} \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \\ ... \\ \epsilon_{N,t} \end{bmatrix}, \quad (26)$$

[9] In particular, the approach of Sims and Zha (1998) achieves conceptual and computational gains by (i) working on the *structural* representation of the VAR, in which the matrix of the errors is diagonal (an identity matrix in their normalization scheme) and (ii) allowing independence across the coefficients belonging to different *structural* equations, which amounts to the prior variance of the coefficients being block-diagonal, which is desirable as it breaks the unreasonable symmetry across equations implied by the conjugate N-IW prior. These two ingredients ensure that the posterior variance matrix has a block-diagonal structure, and therefore achieves computational gains of order N^2. However, such a strategy still implies that the beliefs about the *reduced form* coefficients are correlated across equations in a way that depends on the covariance of the reduced form errors of the model, and gains are not attainable if one wants to impose an asymmetric prior on these *reduced form* coefficients, as explained in section 5.2 of their paper.

where $a^*_{j,i}$ denotes the generic element of the matrix A^{-1} which is available under knowledge of A. We will also denote by $\pi^{(i)}$ the vector of coefficients for equation i contained in row i of Π, for the intercept and coefficients on lagged y_t. The VAR can be written as:

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

with the generic equation for variable j:

$$y_{j,t} - (a^*_{j,1}\lambda_{1,t}^{0.5}\epsilon_{1,t} + ... + a^*_{j,,j-1}\lambda_{j-1,t}^{0.5}\epsilon_{j-1,t}) = \pi_j^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{j,l}^{(i)} y_{i,t-l} + \lambda_{j,t}^{0.5}\epsilon_{j,t}. \tag{27}$$

Consider estimating these equations in order from $j = 1$ to $j = N$. When estimating the generic equation j the term on the left hand side in (27) 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:

$$y^*_{j,t} = y_{j,t} - (a^*_{j,1}\lambda_{1,t}^{0.5}\epsilon_{1,t} + ... + a^*_{j,,j-1}\lambda_{j-1,t}^{0.5}\epsilon_{j-1,t}), \tag{28}$$

and equation (27) becomes a standard generalized linear regression model for the variables in equation (28), with independent Gaussian disturbances with mean 0 and variance $\lambda_{j,t}$. The distribution (10) can be factorized as:

$$\begin{aligned}
p(\Pi|A,\Lambda_T,y) &= p(\pi^{(N)}|\pi^{(N-1)},\pi^{(N-2)},\ldots,\pi^{(1)},A,\Lambda_T,y) \\
&\quad \times p(\pi^{(N-1)}|\pi^{(N-2)},\ldots,\pi^{(1)},A,\Lambda_T,y) \\
&\quad \vdots \\
&\quad \times p(\pi^{(1)}|A,\Lambda_T,y),
\end{aligned} \tag{29}$$

with generic element:

$$\begin{aligned}
p(\pi^{(j)}|\pi^{(j-1)},\pi^{(j-2)},\ldots,\pi^{(1)},A,\Lambda_T,y) &= p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}},A,\Phi,\Lambda_T,y) \\
&\propto p(y|\Pi^{\{j\}},\Pi^{\{1:j-1\}},A,\Lambda_T)p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}),
\end{aligned}$$

where $\Pi^{\{j\}} = \pi^{(j)'}$ denotes the (transposed) j-th row of the matrix Π, and $\Pi^{\{1:j-1\}}$ all of the previous $1, ..., j-1$ rows (transposed). The term $p(y|\Pi^{\{j\}},\Pi^{\{1:j-1\}},A,\Lambda_T)$ is the likelihood of

equation j which coincides with the likelihood of the general linear regression model in (27). The term $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}})$ is the prior on the coefficients of the j-th equation, conditionally on the previous equations. The moments of $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}})$ can be found recursively from the joint prior (7) using $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}) = p(\Pi^{\{j\}}, \Pi^{\{1:j-1\}})/p(\Pi^{\{1:j-1\}})$.

It follows that using the factorization in (29) together with the model in (27) allows one to draw the coefficients of the matrix Π in separate blocks $\Pi^{\{j\}}$ which can be obtained from:

$$\Pi^{\{j\}}|\Pi^{\{1:j-1\}}, A, \Lambda_T, y \sim N(\bar{\mu}_{\Pi^{\{j\}}}, \overline{\Omega}_{\Pi^{\{j\}}}) \tag{30}$$

with

$$\bar{\mu}_{\Pi^{\{j\}}} = \overline{\Omega}_{\Pi^{\{j\}}} \left\{ \underline{\Omega}^{-1}_{\Pi^{\{j\}}} \underline{\mu}_{\Pi^{\{j\}}} + \sum_{t=1}^{T} X_{j,t} \lambda^{-1}_{j,t} y^{*\prime}_{j,t} \right\} \tag{31}$$

$$\overline{\Omega}^{-1}_{\Pi^{\{j\}}} = \underline{\Omega}^{-1}_{\Pi^{\{j\}}} + \sum_{t=1}^{T} X_{j,t} \lambda^{-1}_{j,t} X'_{j,t}, \tag{32}$$

where $y^*_{j,t}$ is defined in (28) 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 (i.e. $p(\Pi^{\{j\}}|\Pi^{\{1:j-1\}}) = p(\Pi^{\{j\}})$). This assumption is frequent in the literature,[10] but can be easily relaxed and we discuss how to do so below. Therefore, the joint posterior distribution of Π can be simulated recursively in separate blocks $\Pi^{\{1\}}, \Pi^{\{2\}}|\Pi^{\{1\}}, \Pi^{\{3\}}|\Pi^{\{1:2\}}, ..., \Pi^{\{N\}}|\Pi^{\{1:N-1\}}$ using (30). Note that this amounts to simple Monte Carlo simulation which will produce draws numerically identical to those that would be obtained using system-wide estimation, meaning that any difference in the simulated posterior draws will be due to random variation (which eventually vanishes) and rounding numerical errors.

The dimension of the matrix $\overline{\Omega}^{-1}_{\Pi^{\{j\}}}$ in (32) 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 estimation algorithm is $O(N^4)$. This is considerably smaller than the complexity of $O(N^6)$ implied by the standard estimation algorithm, with a gain of N^2. For a model with 20

[10] Some widely used priors within the independent N-IW paradigm involve prior correlations among coefficients of the same equations, but not across equations. These include the sum of coefficients and unit root prior proposed by Sims (1993) and Sims and Zha (1998). As we already mentioned, the conjugate prior for a homoskedastic VAR in (24) does impose prior dependence across equations, but for this case an algorithm of computational complexity $O(N^3)$ is already available.

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 triangularizing the system is not exploited, and all computations are carried out using the whole vectorized system. In our algorithm, instead, the triangularization allows one to estimate equations which at most contain $Np + 1$ regressors, and the correlation among the different equations typical of SUR models is implicitly accounted for by the triangularization scheme.

While prior independence across equations is typical in the most common priors elicited in the literature, there might be cases in which a researcher wishes to specify priors which feature correlations across coefficients belonging to different equations. Examples in which a correlation across coefficients of different equations might be expected a priori include rational expectations — present-value models such as the expectation theory of the term structure of interest rates, the uncovered interest rate parity, and the permanent income hypothesis (see, e.g., Campbell and Shiller 1987). For this case, the general form of the posterior can be obtained easily using a similar triangularization argument on the joint prior distribution, and equation (30) generalizes to:

$$\Pi^{\{j\}}|\Pi^{\{1:j-1\}}, A, \Lambda_T, y \sim N(\bar{\mu}_{\Pi^{\{j|1:j-1\}}}, \overline{\Omega}_{\Pi^{\{j|1:j-1\}}}) \tag{33}$$

with

$$\bar{\mu}_{\Pi^{\{j|1:j-1\}}} = \overline{\Omega}_{\Pi^{\{j|1:j-1\}}} \left\{ \sum_{t=1}^{T} X_{j,t} \lambda_{j,t}^{-1} y_{j,t}^{*\prime} + \underline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} \underline{\mu}_{\Pi^{\{j|1:j-1\}}} \right\} \tag{34}$$

$$\overline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} = \underline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} + \sum_{t=1}^{T} X_{j,t} \lambda_{j,t}^{-1} X_{j,t}^{\prime}, \tag{35}$$

where $\underline{\mu}_{\Pi^{\{j|1:j-1\}}}$ and $\underline{\Omega}_{\Pi^{\{j|1:j-1\}}}$ are the moments of $\Pi^{\{j\}}|\Pi^{\{1:j-1\}} \sim N(\underline{\mu}_{\Pi^{\{j|1:j-1\}}}, \underline{\Omega}_{\Pi^{\{j|1:j-1\}}})$, i.e. the conditional priors (for equation j conditional on all of the previous equations) implied by the joint prior specification. The conditional prior moments can be obtained recursively using (19) and standard results on multivariate Gaussian distributions:

$$\underline{\mu}_{\Pi^{\{j|1:j-1\}}} = \underline{\mu}_{\Pi^{\{j\}}} + \underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}} \underline{\Omega}^{-1}_{\Pi^{\{[1:j-1][1:j-1]\}}} (\Pi^{\{1:j-1\}} - \underline{\mu}_{\Pi\{1:j-1\}}), \tag{36}$$

$$\underline{\Omega}_{\Pi^{\{j|1:j-1\}}} = \underline{\Omega}_{\Pi^{\{j\}}} - \underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}} \underline{\Omega}^{-1}_{\Pi^{\{[1:j-1][1:j-1]\}}} \underline{\Omega}^{\prime}_{\Pi^{\{[j][1:j-1]\}}} \tag{37}$$

where $\underline{\Omega}_{\Pi^{\{j\}}}$ denotes the block of $\underline{\Omega}_{\Pi}$ corresponding to equation j, $\underline{\Omega}_{\Pi^{\{[1:j-1][1:j-1]\}}}$ denotes all the blocks on the main block-diagonal, north-west of $\underline{\Omega}_{\Pi^{\{j\}}}$, and $\underline{\Omega}_{\Pi^{\{[j][1:j-1]\}}}$ denotes all the blocks to the left of $\underline{\Omega}_{\Pi^{\{j\}}}$. The computational cost of deriving these conditional prior moments is negligible as they need to be computed only once outside the main MCMC

sampler. Clearly in case of a prior independent across equations, $\underline{\Omega}_{\Pi\{[j][1:j-1]\}}$ is a zero matrix and these expressions simplify to $\underline{\mu}_{\Pi\{j|1:j-1\}} = \underline{\mu}_{\Pi\{j\}}$ and $\underline{\Omega}_{\Pi\{j|1:j-1\}} = \underline{\Omega}_{\Pi\{j\}}$, yielding (31) and (32).

Most of the priors typically used in the literature are entirely compatible with the algorithm described above, including the Minnesota prior (possibly with cross-variable shrinkage), the Sims and Zha (1998) priors (including the sum of coefficients and dummy initial observation priors), the steady-state prior of Villani (2009), the long run prior of Giannone, et al. (2016), and theory-based priors such as those of Ingram and Whiteman (1989) and Del Negro and Schorfheide (2004).

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 (30), with prior mean and variance given by formulas (31) and (32) (or (33), (34), and (35) in case of prior dependence), with the only difference being that the subscript t would be omitted from the volatility terms $\lambda_{j,t}$. 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. For the homoskedastic case Waggoner and Zha (2003) proposed an efficient Gibbs sampler which is also based on an equation-by-equation approach, and Koop, et al. (2016) proposed to use the method of compression to achieve computational gains. However these approaches are grounded on the Sims and Zha (1998) prior specification, and as such they cannot handle the case of asymmetric priors for the reduced form parameters.

3.1 The role of variable ordering

It is worth stressing that expression (26) 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. Once the posterior draws have been obtained at the end of the MCMC simulations, any identification scheme can be applied to perform structural analysis, including different Cholesky orderings, sign restrictions, and long run restrictions. The implied structural form matrices can be readily obtained from the reduced form posteriors.

Moreover, note that under knowledge of Σ_t, the ordering of the variables in the system does not change the conditional posterior distribution of the reduced form coefficients, so changing the order of the variables is inconsequential to the resulting conditional posterior of Π. However, note that the latter statement relates to drawing from the conditional posterior of the conditional mean parameters, when Σ_t belongs to the conditioning set. One needs also

to keep in mind that the joint distribution of the system might be affected by the ordering of the variables in the system due to an entirely different reason: the use of the diagonalization (3) typically used for Σ_t in stochastic volatility models. Since priors are elicited separately for A and Λ_t, the implied prior of Σ_t will change if one changes the equation ordering, and therefore different orderings would result in different prior specifications and then potentially different joint posteriors. This problem is not a feature of our algorithm, but rather it is inherent to all models using the diagonalization (3).

As noted by Primiceri (2005), this problem will be mitigated in the case (as the one considered in this paper) in which the covariances A do not vary with time, because the likelihood will quickly dominate the prior as the sample size increases. This problem can entirely be avoided by eliciting a prior on the whole matrix Σ_t rather than proceeding with its diagonalization; for example Shin and Zhong (2016) do so using the multivariate stochastic volatility specification of Philipov and Glickman (2006). Finally, it is important to note that this problem is entirely absent in the case of the homoskedastic model, since in such a case one can easily specify the prior directly on the whole matrix Σ.

3.2 MCMC samplers

We summarize the steps involved in the MCMC samplers for the BVAR with stochastic volatility and for a BVAR with asymmetric priors (generally, a BVAR with independent Normal-Wishart priors) highlighting how all the existing 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.

3.2.1 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 Carter and Kohn (1994).

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 (30) and the factorization (29).

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). Also, with independence, the IW step for drawing Φ would be replaced by a step to loop over the N variables to draw each element on the diagonal Φ.

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.

3.2.2 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 (30) and the factorization (29).

Step 2: Draw the matrix Σ conditional on the data and Π, using the conditional (IW) distribution for the posterior given in (22).

Note that the only difference between this algorithm and the standard algorithm used, e.g., in Karlsson (2013) 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.

4 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).

4.1 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 (2016) (MN dataset), at monthly frequency, from January 1960 to December 2014. The data are transformed as in McCracken and Ng (2016) 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 widely followed, aggregate 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.[11] 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).

[11] 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.

Figure 2 further illustrates the computational gains arising from the use of the triangular algorithm. The top panel shows the computational time (on a 3.5 GHz Intel Core i7) needed to perform 10 draws as a function of the size of the cross section using the triangular algorithm (blue line) and the system-wide algorithm (red 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.

Indeed, we do not extend this comparison to $N = 125$, which is the size used in the empirical application we present below in Section 5, because for a model of this size the system-wide algorithm would be extremely computationally demanding: a scalar number stored in double-precision floating-point format requires 8 bytes, and for a system $N = 125$ the size of the covariance matrix of the coefficients is of dimension 203250, which would require about 330 GB of RAM ($203250^2 \times 8/10^9$).[12]

4.2 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 of 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.

[12]For a reference, consider that most desktops nowadays have either 8 or 16 GB of RAM.

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 our 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 our 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. In the figure, both algorithms are initialized at the same value, given by the prior means and variances of the coefficients.

Since these gains increase nonlinearly with the system size, we conclude that, for forecasting or structural analysis with medium and large BVARs, our estimation method based on 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.

5 A large structural VAR with drifting volatilities

In this Section we illustrate how our estimation method based on the triangular algorithm can be used to estimate a very large BVAR with drifting volatilities and asymmetric priors. We consider a VAR with 125 variables, which includes all of the variables considered by McCracken and Ng (2016) 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, so the prior is asymmetric and could not be cast in the form (24). Furthermore, all of the errors feature stochastic volatility. The appendix provides additional details on the priors.

The total number of objects to be estimated is extremely large: 203250 mean coefficients, 7750 covariance coefficients, 125 latent states (each of length T), and 7875 covariances of the states. Despite the huge dimension of the system, our estimation algorithm can produce 5000 draws (after 500 of burning in) in just above 7 hours on a 3.5 GHz Intel Core i7.

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.

5.1 Volatilities

Figures 7 and 8 graph the estimated volatilities for, respectively, slow and fast variables, where the classification of fast and slow is as close as possible to Bernanke, Boivin and Eliasz (2005, BBE). 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 45% of overall variance, and the first three 73%, confirming that commonality is indeed present but idiosyncratic movements also matter (as in the GFSV specification of Carriero, et al. (2016) and the factor volatility specification of Carriero, et al. (2017)).

5.2 Impulse responses

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.

The impulse responses present patterns in line with economic theory, 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 actual responses that are proportional to the actual size of the shock) over time.

Stochastic volatility would also matter for variance decompositions, omitted here in the interest of brevity.

5.3 The factor structure of time varying volatilities

The joint posterior of the model coefficients also provides us with an estimate of the matrix Φ, which describes the covariance structure of the shocks hitting the panel of volatilities. It is an interesting question to look at how these shocks are related to different types of variables in the system. To do so, we have performed a Principal Component (PC) analysis on the posterior mean of Φ, and we have collected the loadings associated with the first 5 principal components in Figure 11. The figure shows that the panel of volatilities is largely driven by two shocks, the first impacting the volatilities of all variables in the panel and explaining the largest part (73%) of the total variation in the volatilities, while the second shock, explaining a further 19% of variation, is mostly impacting financial variables. A third shock accounts for only 2.6% of residual variation. This result shows that movements in the volatilities of macroeconomic variables are largely driven by two main shocks, which can be interpreted as representing macroeconomic uncertainty and financial uncertainty. This argument is further developed in Carriero, et al. (2017).

Furthermore, it is possible to perform PC analysis on the volatility states. Conceptually,

these represent the static factors corresponding to the two dynamic factors highlighted above. The PC analysis results are displayed in Figures 12 and 13 and show that 3 factors are sufficient to explain about 73% of the total variation in the volatilities, with the first factor accounting for about 45%, the second for about 15%, and the third for about 12%. Interestingly, such factors can be clearly linked to specific groups of macroeconomic data. A first factor (blue in Figure 13) mainly loads on real activity variables (see the groupings in Figure 12), and its fluctuations induce the pattern in all the variables included in this group (these are the first 52 volatilities in Figure 7). A second factor (red in Figure 13) mainly loads on prices and monetary aggregates (see the groupings in Figure 12), and its fluctuations induce the pattern in all the variables included in this group (these are the volatilities in positions 53 to 72 and 105 to 118 in Figures 7 and 8). A third factor (green in Figure 13) mainly loads on the FFR and other interest rates and financial variables (see groupings in Figure 12) and its fluctuations induce the pattern in this group (these are the volatilities 73 to 104 in Figure 8).

The results of this analysis open the way to build a model in which volatilities follow a factor structure, with different factors representing different types of uncertainty, a strategy which we pursue in Carriero, Clark and Marcellino (2017).

6 The role of model size and stochastic volatility for forecasting

The previous section showed that a large BVAR with time varying volatility can offer several insights regarding the impulse propagation mechanism of monetary policy shocks and the underlying shocks driving the volatilities of all variables. Besides structural analysis, models are commonly used to forecast the future behavior of macroeconomic time series, and they are compared on the basis of their forecasting performance. In this section we investigate the implications that cross-sectional size and time varying volatilities have on the out-of-sample forecasting performance of a VAR.[13]

We perform a recursive pseudo out-of-sample forecasting exercise to study the effects that time variation in volatility and the use of a large information set have on the precision of density and point forecasts of macroeconomic variables. The out-of-sample exercise is

[13] As noted by Diebold (2015) pseudo-out-of-sample forecasting exercise are not superior to several others model comparison techniques, notably F-tests and posterior odds, and are actually less powerful. However, performing posterior odds analysis presents problems in the case at hand because for the independent N-IW prior used in this paper the marginal likelihood is not available in closed form and its computation would require an extremely demanding Monte Carlo integration.

performed recursively, starting with the estimation sample 1960:3 to 1970:2 (ten years of monthly data) and ending with the estimation sample 1960:3 to 2014:5. We compute forecasts up to 12 steps ahead; therefore the forecasting samples range from 1970:3-1971:2 to 2014:6-2015:5, for a total of 531 sets of 12-step ahead forecasts.

We consider four models. The first model is a small homoskedastic VAR including the growth rate of industrial production ($\Delta \ln IP$), the inflation rate based on consumption expenditures ($\Delta \ln PCEPI$) and the effective Federal Funds Rate (FFR). The second model is also a homoskedastic VAR, but includes the 20 macroeconomic variables identified by a star in Table 1 (these are the same variables used in the numerical comparison of Section 4). As similar models have been shown to be very competitive in forecasting in papers such as Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015) and Koop (2013) we set this as our benchmark; namely, we will provide results relative to the performance of this model. The third model instead is still based on a tri-variate specification, but it allows for time variation in volatilities. Also small models of this type have received support in the literature in terms of their forecasting performance; see, e.g., Clark (2011), Clark and Ravazzolo (2015), Cogley, Morozov, and Sargent (2005), and D'Agostino, Gambetti and Giannone (2013). Moreover, models of this scale have been used in the structural analyses of Cogley and Sargent (2005) and Primiceri (2005). The fourth model includes both time variation in the volatilities and a large (20 variables) information set, thereby using both the ingredients that seem to be important to improve density and point forecasts. This model can be rather easily estimated using the approach proposed in this paper.[14]

A priori, we expect the inclusion of time variation in volatilities to improve density forecasts via a better modeling of error variances, while the use of a larger dataset should improve point forecasts via a better specification of the conditional means. However, this is not the whole story, as there are also interaction effects: a better point forecast should improve the density forecast as well, by centering the predictive density around a more reliable mean, and time varying volatilities should improve the point forecasts — especially at longer horizons — because the heteroskedastic model will provide more efficient estimates (through a GLS argument) and therefore a better characterization of the predictive densities, with the predictive means gradually deviating from their homoskedastic counterparts as the predictive densities cumulate nonlinearly with the forecast horizon.

Indeed, this is precisely the pattern we find in the data. Figure 14 displays the Root

[14]Naturally, larger forecasting models could be also used, but as we recursively repeat the exercise many times this becomes computationally demanding.

Mean Squared Forecast Error (RMSFE) relative (ratio) to the benchmark (the 20 variables homoskedastic VAR), so that a value below 1 denotes a model outperforming the benchmark. The large homoskedastic model outperforms the small homoskedastic model for all variables at all horizons, suggesting that the inclusion of more data does improve the specification of the conditional means and therefore the point forecasts. The inclusion of time variation in volatilities consistently improves the performance of the small model, and for the FFR it also outperforms the benchmark at long horizons. However, the small heteroskedastic model is still largely dominated by the benchmark at short forecast horizons. The model with both time varying volatilities and a large cross section instead provides systematically better point forecasts than the benchmark (and than the other models), with the only exception of inflation for the 1, 2, and 3 step-ahead horizons.

Figure 15 presents results for density forecasts, based on the average log scores. The Figure displays the average log scores relative (difference) to the benchmark (the 20 variables homoskedastic VAR), so that a value above 0 denotes a model outperforming the benchmark. Both homoskedastic specifications perform quite poorly in density forecasting, while the heteroskedastic specifications can achieve very high gains. Moreover, the large heteroskedastic system consistently outperforms the small heteroskedastic system. In combination with the results presented for point forecasts, this result suggests that while both the heteroskedastic models provide a better assessment of the overall uncertainty around the forecasts, the model based on the large cross section centers such uncertainty around a more reliable mean, thereby obtaining further gains in predictive accuracy.

So far the discussion has been focused only on the three variables industrial production, inflation, and Federal Funds rate. For the larger specifications (the BVARs with 20 variables) is of course possible to compare forecasts for all the variables included in the cross section. Results of this comparison are displayed in Figure 16 (for point forecasts) and Figure 17 (for density forecasts). In these graphs each subplot corresponds to a different variable.

In all of the subplots in Figure 16 the x axes measure the RMSFE obtained by the large BVAR when we allow for stochastic volatility, while the y axes measure the same loss function (RMSFE) obtained by the homoskedastic specification. Each point corresponds to a different forecast horizon, and when a point is *above* the 45 degree line this shows that the RMSFE of the heteroskedastic specification is smaller, indicating that the inclusion of variation in the volatility improved point forecasting performance. As is clear in the graph, in several instances the models produce similar point forecasts. However, as the forecast horizon increases (which can be indirectly inferred from the graph as in general higher RMSFE correspond to longer forecast horizons) the specification with variation in

the volatilities tends to outperform the homoskedastic version of the model. The mechanism at play is as follows: the heteroskedastic model provides more efficient estimates and a therefore a better characterization of the predictive densities, while the homoskedastic model is misspecified and therefore provides an inferior characterization of the predictive densities. At short forecast horizons this does not have much effect on point forecasts, but as the forecast horizon increases, the predictive densities cumulate nonlinearly and therefore the misspecification of the homoskedastic model increasingly reduces the relative accuracy of the point forecasts.

We now turn to density forecasts, which are described in Figure 17. In the subplots in Figure 17 the x axes measure the (log) density score obtained by the large BVAR when we allow for stochastic volatility, while the y axes measure the same gain function (score) obtained by the homoskedastic specification. Each point corresponds to a different forecast horizon, and when a point is *below* the 45 degree line this shows that the score of the heteroskedastic specification is larger, indicating that the inclusion of variation in the volatility improved density forecasting performance. In Figure 17 the improvement coming from the introduction of time variation in the volatilities is striking and it is common to all variables, with only a handful of exceptions. Clearly, stochastic volatility improves the overall assessment of uncertainty with respect to the homoskedastic model, and it does so both directly, by simply using the correct variance around the point estimates, and indirectly, by centering the densities towards improved point forecasts (as documented in Figure 16).

To summarize, the joint use of time varying volatilities and a large cross-section produces forecasting gains larger than those obtained by using either of this two ingredients separately.

7 Conclusions

Recent research has shown that a reliable Vector Autoregressive model (VAR) for forecasting and structural analysis of macroeconomic data requires a large set of variables and modeling time variation in their volatilities. Yet, there are no papers jointly allowing for stochastic volatilities and large datasets, due to computational complexity. Moreover, homoskedastic VAR models for large datasets so far restrict substantially the allowed prior distributions on the parameters.

In this paper we have proposed a new estimation method for large VARs with possibly asymmetric priors and drifting volatilities. The method is based on a straightforward triangularization of the system, and it is very simple to implement. Indeed, if a researcher already has algorithms to produce draws from a VAR with an independent N-IW prior and

stochastic volatility, only the step in which the conditional mean parameters are drawn needs to be modified, which can be easily done with a few lines of code.

The algorithm ensures computational gains of order N^2 with respect to the traditional algorithm used to estimate VARs with time varying volatilities, and because of this it is possible to achieve much better mixing and convergence properties compared to existing algorithms and substantial computational gains. This makes estimation of this type of model doable regardless of the dimension of the system. 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, especially those involving large datasets.

Moreover, our approach makes viable the estimation of models with independent N-IW priors (as well as Normal-diffuse priors) of any model size. Since the independent N-IW prior is much more flexible than the conjugate N-IW prior, we argue that it should be preferred in most situations, including some in which the model is homoskedastic. The conjugate N-IW prior imposes restrictions on the prior covariance matrix of the coefficients which can be in many instances undesirable, since it implies that the prior precision has to be the same (up to a scaling factor) in all equations, and that coefficients belonging to different equations have to be correlated with a correlation structure proportional to that of the error variance.

We have presented a numerical example to show that the new and old algorithms lead to draws from the same posterior distribution, apart from random deviations, and so for example also lead to the same impulse response functions and forecasts. The only, but main, difference is computational time and efficiency.

Then, we have illustrated the empirical application of the new estimation method by studying the effects of a monetary policy shock in a large Vector Autoregression with stochastic volatilities, finding interesting patterns in the latter, in the response functions, and in the time-varying size of the shock.

Finally, we have shown how, jointly, the inclusion of time varying volatilities and the use of a large data-set improve point and density forecasts for macroeconomic and financial variables, with gains that are larger than what would be obtained by using these two ingredients separately.

In closing we want to highlight two caveats. First, while the independent N-IW prior avoids putting on the data the straightjacket that the conjugate N-IW does, the computation of the marginal likelihood is not as simple, while for the conjugate N-IW prior the marginal likelihood is available in closed form (for homoskedastic models). Second, while the model with stochastic volatility does produce dramatically superior density forecasts than its homoskedastic counterpart, some work is still needed to improve the density forecasts in

the exact periods a large swing in volatilities takes place. Both these issues require further research.

8 Appendix

8.1 Specifics on priors

In this section we discuss in detail the priors used for the BVARs estimated in the paper. The priors for the coefficients blocks of the model are as follows:

$$\text{vec}(\Pi) \sim N(\text{vec}(\underline{\mu}_{\Pi}), \underline{\Omega}_{\Pi}); \tag{38}$$

$$A \sim N(\underline{\mu}_{A}, \underline{\Omega}_{A}); \tag{39}$$

$$\Phi \sim IW(\underline{d}_{\Phi} \cdot \underline{\Phi}, \underline{d}_{\Phi}). \tag{40}$$

The prior moments of the VAR coefficients $\underline{\mu}_{\Pi}$ and $\underline{\Omega}_{\Pi}$ are specified along the lines of the Minnesota prior beliefs. In particular, for the conditional mean coefficients, we set a prior mean of 0 for the intercepts Π_0 and for all the coefficient matrices in the matrix polynomial $\Pi(L)$ for $L = 2, ..., p$. The lag-1 coefficient matrix Π_1 is set to a diagonal matrix with diagonal elements being either 1 or 0 depending on the degree of persistence (high or low) of the series included in the estimation. The prior variances are specified as in Litterman's (1979) original implementation of the Minnesota prior, which includes cross-variable shrinkage. In particular we set $\underline{\Omega}_{\Pi}$ such that:

$$Var(\Pi_l^{(ij)}) = \frac{\lambda_1 \lambda_2}{l^{\lambda_3}} \frac{\sigma_i^2}{\sigma_j^2}, \; l = 1, ..., p \tag{41}$$

where $\Pi_l^{(ij)}$ denotes the element in row i and column j of the matrix Π_l. For the intercepts we elicit an uninformative prior by setting the prior variance equal to 100. The parameter λ_1 measures the overall tightness of the prior and is set to 0.05. The parameter λ_2 implements additional shrinkage on lags of other variables than for lags of the dependent variable and we set it to 0.5. The parameter λ_3 measures determines the rate at which the prior variance decreases with increasing lag length and is set to 2 (quadratic decay). To set the scale parameters σ_i^2 we follow common practice (see e.g. Litterman, 1986; Sims and Zha, 1998) and set it equal to the variance of the residuals from a univariate autoregressive model.

The matrix A collects the covariances of the errors. We set each individual element of this matrix to be a-priori normally distributed with means collected in the vector $\underline{\mu}_{A}$ and variances collected in the vector $\underline{\Omega}_{A}$. The prior means $\underline{\mu}_{A}$ are all set to 0 while the prior variance is uninformative, set to a diagonal matrix with diagonal elements 10^6 (this implements a virtually flat prior on these coefficients).

The matrix Φ is the variance matrix of the innovations to the volatilities. It is set a-priori to an Inverse Wishart distribution with scale $\underline{d}_\Phi \cdot \underline{\Phi}$ and $\underline{d}_\Phi$ degrees of freedom. The degrees of freedom are set to $N + 2$ which provides the least informative proper prior. The scale matrix is set to an identity matrix.

8.2 Volatility estimation

Our treatment of volatility draws on Primiceri's (2005) implementation of the Kim, Shephard, and Chib (1998) algorithm (hereafter, KSC algorithm). As indicated above, v_t denotes the reduced form residuals of the VAR and $\tilde{v}_t = \tilde{A}v_t$ are the rescaled residuals, which obey equation (4). We further define $v^*_{j,t} = \ln(\tilde{v}^2_{j,t} + \bar{c})$, where $\bar{c}$ denotes an offset constant used in the KSC algorithm. With this notation, we can establish the measurement equation of a state-space system with non-Gaussian errors:

$$v^*_{j,t} = \ln \lambda_{j,t} + \ln \epsilon^2_{j,t},\ j = 1, \ldots, N. \tag{42}$$

The transition equations are given by (5). In the equations above $\ln \epsilon^2_{j,t}$ is not Gaussian, but $\epsilon_{j,t}$ is a Gaussian process with unit variance, and with this setup we can use the mixture of normals approximation of KSC to estimate volatility with a Gibbs sampler, first drawing the states of the mixture and then drawing volatility conditional on the states. Primiceri (2005) and Del Negro and Primiceri (2015) detail the steps required. Alternatively, if the innovations to volatility are assumed to be uncorrelated (Φ diagonal), one can use the Cogley and Sargent (2005) approach to draw the volatility states.

The prior specification is completed by eliciting a prior for the initial values of the state variables Λ_t, which we set to independent Gaussian distributions with mean 0 and variance 100.

References

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

[2] Bernanke, B., Boivin, J., and Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach, Quarterly Journal of Economics 120, 387-422.

[3] Campbell, J., and Shiller, R., 1987. Cointegration and tests of present value models, Journal of Political Economy 95, 1062-1088.

[4] Carriero, A., Clark, T., and Marcellino, M., 2015. Bayesian VARs: specification choices and forecast accuracy, Journal of Applied Econometrics 30, 46-73.

[5] Carriero, A., Clark, T., and Marcellino, M., 2016. Common drifting volatility in large Bayesian VARs, Journal of Business and Economic Statistics 34, 375-390.

[6] Carriero, A., Clark, T., and Marcellino, M., 2017. Measuring uncertainty and its effects on the economy, Review of Economics and Statistics, forthcoming.

[7] Carter, C., and Kohn, R., 1994. On Gibbs sampling for state space models, Biometrika 81, 541-553.

[8] Chan, J., 2015. Large Bayesian VARs: a flexible Kronecker error covariance structure, manuscript.

[9] Chib, S., and Greenberg, E., 1995. Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models, Journal of Econometrics 68, 339-360.

[10] Clark, T., 2011. Real-time density forecasts from BVARs with stochastic volatility, Journal of Business and Economic Statistics 29, 327-341.

[11] Clark, T., and Ravazzolo, F., 2015. Macroeconomic forecasting performance under alternative specifications of time-varying volatility, Journal of Applied Econometrics 30, 551-575.

[12] Cogley, T., Morozov, S., and Sargent, T., 2005. Bayesian fan charts for U.K. inflation: forecasting and sources of uncertainty in an evolving monetary system, Journal of Economic Dynamics and Control 29, 1893-1925.

[13] 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.

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

[15] Del Negro, M., and Primiceri, G., 2015. Time varying structural vector autoregressions and monetary policy: a corrigendum, Review of Economic Studies 82, 1342-1345.

[16] Del Negro, M., and Schorfheide, F., 2004. Priors from general equilibrium models for VARs, International Economic Review 45, 643-673.

[17] Diebold, F., 2015. Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests, Journal of Business and Economic Statistics 33, 1-9.

[18] Geweke, J., and Whiteman, C., 2006. Bayesian forecasting, In: G. Elliott, C.W.J. Granger, and A. Timmermann, (Eds.), Handbook of Economic Forecasting, Volume 1, 3-80, Elsevier.

[19] Giannone, D., Lenza, M., and Primiceri, G., 2015. Prior selection for vector autoregressions, Review of Economics and Statistics 97, 436-451.

[20] Giannone, D., Lenza, M., and Primiceri, G., 2016. Priors for the long run, CEPR Discussion Paper No. DP11261.

[21] Ingram, B., and Whiteman, C., 1994. Supplanting the 'Minnesota' prior: forecasting macroeconomic time series using real business cycle model priors, Journal of Monetary Economics 34, 497-510.

[22] Jacquier, E., Polson, N., and Rossi, P., 2002. Bayesian analysis of stochastic volatility models, Journal of Business and Economic Statistics 20, 69-87.

[23] Kadiyala, K., and Karlsson, S., 1993. Forecasting with generalized Bayesian vector autoregressions, Journal of Forecasting 12, 365-378.

[24] Kadiyala, K., and Karlsson, S., 1997. Numerical methods for estimation and inference in Bayesian VAR models, Journal of Applied Econometrics 12, 99-132.

[25] Karlsson, S., 2013. Forecasting with Bayesian vector autoregression, In: G. Elliott and A. Timmermann, (Eds.), Handbook of Economic Forecasting, Volume 2, 791-897, Elsevier.

[26] Kim, S., Shephard, N. and Chib, S., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models, Review of Economic Studies 65, 361-393.

[27] Koop, G., 2013. Forecasting with medium and large Bayesian VARs, Journal of Applied Econometrics 28, 177-203.

[28] Koop, G., and Korobilis, D., 2013. Large time-varying parameter VARs, Journal of Econometrics 177, 185-198.

[29] Koop, G., Korobilis, D., and Pettenuzzo, D., 2016. Bayesian compressed vector autoregressions, Journal of Econometrics, forthcoming.

[30] Korobilis, D., and Pettenuzzo, D., 2017. Adaptive Minnesota prior for high-dimensional vector autoregressions, manuscript, Brandeis University.

[31] Litterman, R., 1979. Techniques of forecasting using vector autoregressions, Federal Reserve Bank of Minneapolis Working Paper, no. 115.

[32] Litterman, R., 1986. Forecasting with Bayesian vector autoregressions — five years of experience, Journal of Business and Economic Statistics 4, 25-38.

[33] McCracken, M., and Ng, S., 2016. FRED-MD: a monthly database for macroeconomic research, Journal of Business and Economic Statistics 34, 574-589.

[34] Philipov, A. and Glickman, M., 2006. Multivariate stochastic volatility via Wishart processes, Journal of Business and Economic Statistics 24, 313-328.

[35] Primiceri, G., 2005. Time varying structural vector autoregressions and monetary policy, Review of Economic Studies 72, 821-852.

[36] Rothenberg, T., 1963. A Bayesian analysis of simultaneous equation systems, report 6315, Econometric Institute, Netherlands School of Economics, Rotterdam.

[37] Shin, M., and Zhong, M., 2016. A new approach to identifying the real effects of uncertainty shocks, manuscript.

[38] Sims, C., 1993. A nine-variable probabilistic macroeconomic forecasting model, in J. Stock and M. Watson, (Eds.), Business Cycles, Indicators and Forecasting, University of Chicago Press, 179-212.

[39] Sims, C., and Zha, T., 1998. Bayesian methods for dynamic multivariate models, International Economic Review 39, 949-968.

[40] Villani, M., 2009. Steady-state priors for vector autoregressions, Journal of Applied Econometrics 24, 630-650.

[41] Waggoner, D., and Zha, T., 2003. A Gibbs sampler for structural vector autoregressions, Journal of Economic Dynamics and Control 28, 349-366.

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

Table 1: Variables used in the 125-dimensional VAR with Minnesota prior and stochastic volatility
(a star indicates inclusion in the 20-variable system)

Slow variables

	variable	mnemonic
1	Real Personal Income*	RPI
2	RPI ex. Transfers	W875RX1
3	Real PCE*	DPCERA3M086SBEA
4	Real M&T Sales*	CMRMTSPLx
5	Retail and Food Services Sales	RETAILx
6	IP Index*	INDPRO
7	IP: Final Products and Supplies	IPFPNSS
8	IP: Final Products	IPFINAL
9	IP: Consumer Goods	IPCONGD
10	IP: Durable Consumer Goods	IPDCONGD
11	IP: Nondurable Consumer Goods	IPNCONGD
12	IP: Business Equipment	IPBUSEQ
13	IP: Materials	IPMAT
14	IP: Durable Materials	IPDMAT
15	IP: Nondurable Materials	IPNMAT
16	IP: Manufacturing	IPMANSICS
17	IP: Residential Utilities	IPB51222S
18	IP: Fuels	IPFUELS
19	Capacity Utilization: Manufacturing*	CUMFNS
20	Help-Wanted Index for US Help wanted indx	HWI
21	Help Wanted to Unemployed ratio	HWIURATIO
22	Civilian Labor Force	CLF16OV
23	Civilian Employment	CE16OV
24	Civilian Unemployment Rate*	UNRATE
25	Average Duration of Unemployment	UEMPMEAN
26	Civilians Unemployed <5 Weeks	UEMPLT5
27	Civilians Unemployed 5-14 Weeks	UEMP5TO14
28	Civilians Unemployed >15 Weeks	UEMP15OV
29	Civilians Unemployed 15-26 Weeks	UEMP15T26
30	Civilians Unemployed >27 Weeks	UEMP27OV
31	Initial Claims	CLAIMSx
32	All Employees: Total nonfarm*	PAYEMS
33	All Employees: Goods-Producing	USGOOD
34	All Employees: Mining and Logging	CES1021000001
35	All Employees: Construction	USCONS
36	All Employees: Manufacturing	MANEMP
37	All Employees: Durable goods	DMANEMP
38	All Employees: Nondurable goods	NDMANEMP
39	All Employees: Service Industries	SRVPRD
40	All Employees: TT&U	USTPU
41	All Employees: Wholesale Trade	USWTRADE
42	All Employees: Retail Trade	USTRADE
43	All Employees: Financial Activities	USFIRE
44	All Employees: Government	USGOVT
45	Hours: Goods-Producing*	CES0600000007

46	Overtime Hours: Manufacturing	AWOTMAN
47	Hours: Manufacturing	AWHMAN
48	Total Business Inventories	BUSINVx
49	Inventories to Sales Ratio	ISRATIOx
50	Ave. Hourly Earnings: Goods*	CES0600000008
51	Ave. Hourly Earnings: Construction	CES2000000008
52	Ave. Hourly Earnings: Manufacturing	CES3000000008
53	PPI: Finished Goods*	PPIFGS
54	PPI: Finished Consumer Goods	PPIFCG
55	PPI: Intermediate Materials	PPIITM
56	PPI: Crude Materials	PPICRM
57	Crude Oil Prices: WTI	oilpricex
58	PPI: Commodities*	PPICMM
59	CPI: All Items	CPIAUCSL
60	CPI: Apparel	CPIAPPSL
61	CPI: Transportation	CPITRNSL
62	CPI: Medical Care	CPIMEDSL
63	CPI: Commodities	CUSR0000SAC
64	CPI: Durables	CUUR0000SAD
65	CPI: Services	CUSR0000SAS
66	CPI: All Items Less Food	CPIULFSL
67	CPI: All items less shelter	CUUR0000SA0L2
68	CPI: All items less medical care	CUSR0000SA0L5
69	PCE: Chain-type Price Index*	PCEPI
70	PCE: Durable goods	DDURRG3M086SBEA
71	PCE: Nondurable goods	DNDGRG3M086SBEA
72	PCE: Services	DSERRG3M086SBEA

Fast variables

	variable	mnemonic
73	Effective Federal Funds Rate*	FEDFUNDS
74	Starts: Total*	HOUST
75	Starts: Northeast	HOUSTNE
76	Starts: Midwest	HOUSTMW
77	Starts: South	HOUSTS
78	Starts: West	HOUSTW
79	Orders: Durable Goods	AMDMNOx
80	Unfilled Orders: Durable Goods	AMDMUOx
81	S&P: Composite*	S&P 500
82	S&P: Industrials	S&P: indust
83	S&P: Dividend Yield	S&P div yield
84	S&P: Price-Earnings Ratio	S&P PE ratio
85	Switzerland / U.S. FX Rate	EXSZUSx
86	Japan / U.S. FX Rate	EXJPUSx
87	U.S. / U.K. FX Rate*	EXUSUKx
88	Canada / U.S. FX Rate	EXCAUSx
89	Month AA Comm. Paper Rate CPF3M Comm paper	CP3Mx
90	3-Month T-bill	TB3MS
91	6-Month T-bill	TB6MS

92	1-year T-bond	GS1
93	5-year T-bond	GS5
94	10-year T-bond	GS10
95	Corporate Bond Yield Aaa bond	AAA
96	Corporate Bond Yield Baa bond	BAA
97	CP - FFR spread CP-FF spread	COMPAPFFx
98	3 Mo. - FFR spread 3 mo-FF spread	TB3SMFFM
99	6 Mo. - FFR spread 6 mo-FF spread	TB6SMFFM
100	1 yr. - FFR spread 1 yr-FF spread*	T1YFFM
101	5 yr. - FFR spread 5 yr-FF spread	T5YFFM
102	10 yr. - FFR spread 10 yr-FF spread*	T10YFFM
103	Aaa - FFR spread Aaa-FF spread	AAAFFM
104	Baa - FFR spread Baa-FF spread*	BAAFFM
105	Money Stock	M1SL
106	Money Stock	M2SL
107	Real M2 Money Stock	M2REAL
108	St. Louis Adjusted Monetary Base	AMBSL
109	Total Reserves	TOTRESNS
110	Nonborrowed Reserves	NONBORRES
111	Commercial and Industrial Loans	BUSLOANS
112	Real Estate Loans	REALLN
113	Total Nonrevolving Credit	NONREVSL
114	Credit to PI ratio	CONSPI
115	MZM Money Stock	MZMSL
116	Consumer Motor Vehicle Loans	DTCOLNVHFNM
117	Total Consumer Loans and Leases	DTCTHFNM
118	Securities in Bank Credit	INVEST
119	ISM Manufacturing: Production	NAPMPI
120	ISM Manufacturing: Employment	NAPMEI
121	ISM: PMI Composite Index	NAPM
122	ISM: New Orders Index*	NAPMNOI
123	ISM: Supplier Deliveries Index	NAPMSDI
124	ISM: Inventories Index	NAPMII
125	ISM Manufacturing: Prices	NAPMPRI

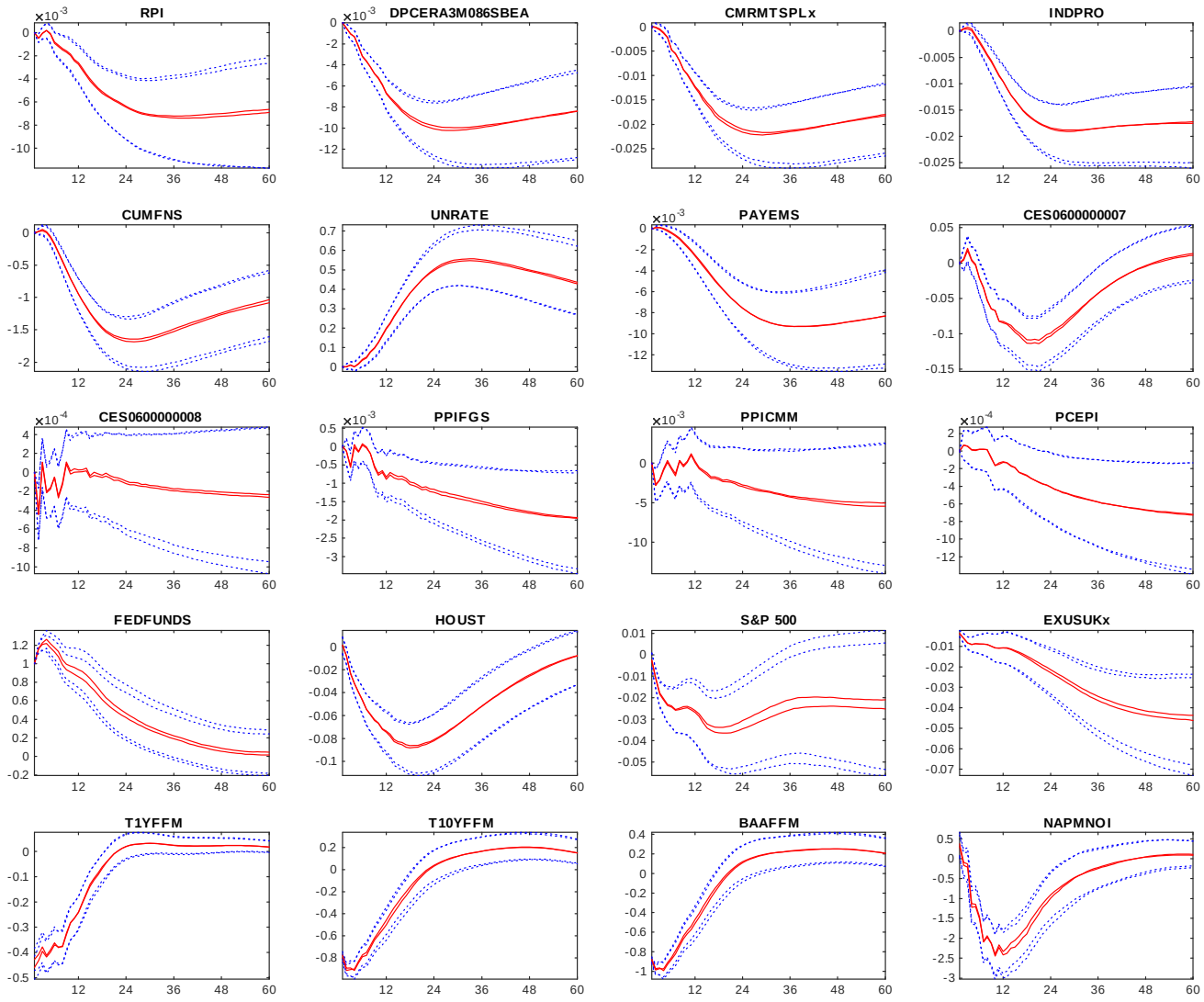

Figure 1: Impulse responses to a Federal Funds Rate shock, estimated under the system-wide and triangular algorithms. For both algorithms the red solid line represents the median response, and the dotted blue lines represent the 16% and 84% quantiles. See Table 1 for a description of the variables.

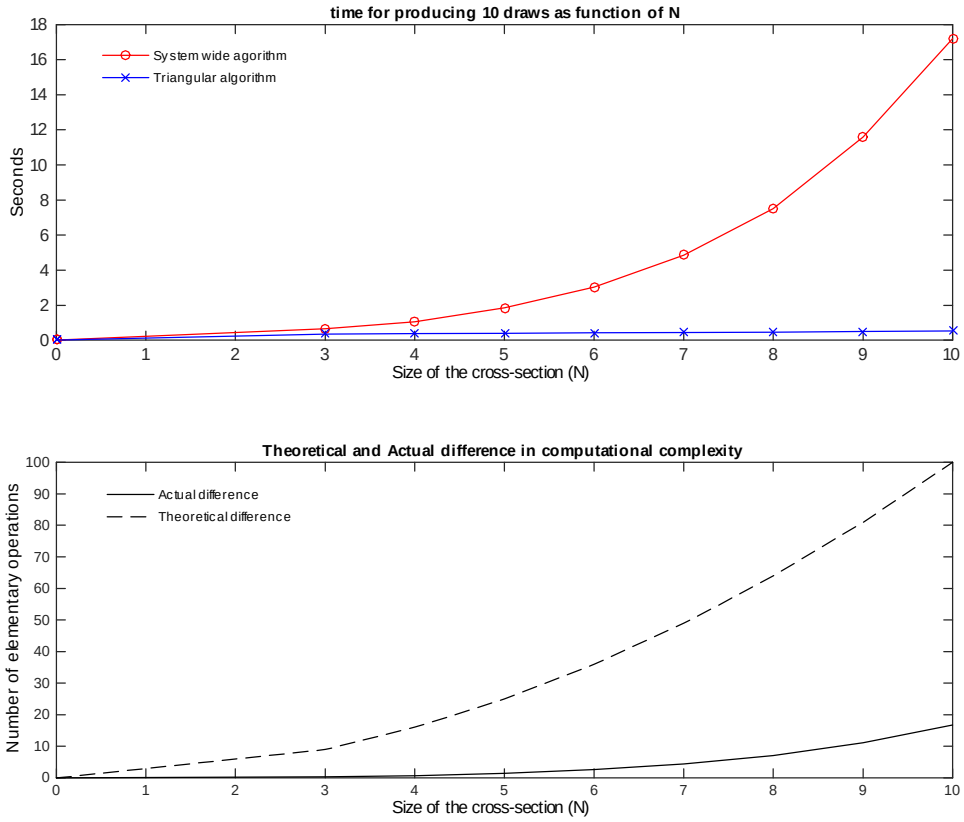

Figure 2: Computational time and complexity of the alternative algorithms for a cross section of less than 10 variables. Computational times are the average time (over 10 independent chains) required to draw 10 draws on a 3.5 GHz Intel Core i7.

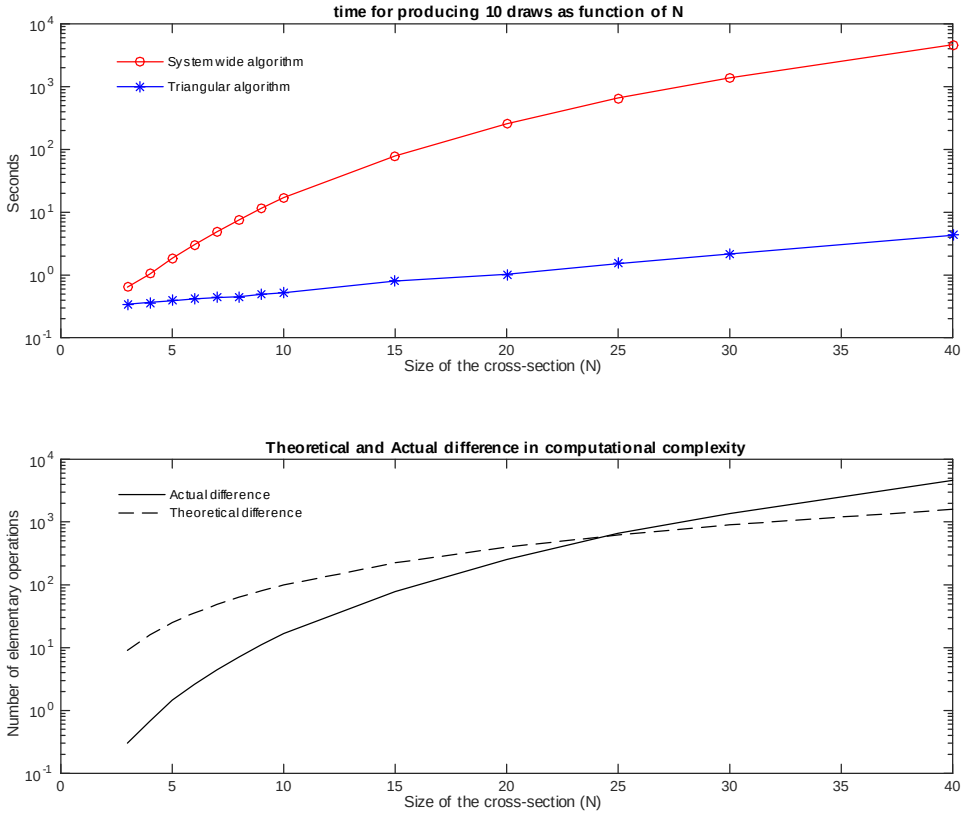

Figure 3: Computational time and complexity of the alternative algorithms for a cross section of more than 10 variables, y-axes are in logarithmic scale. Computational times are the average time (over 10 independent chains) required to draw 10 draws on a 3.5 GHz Intel Core i7.

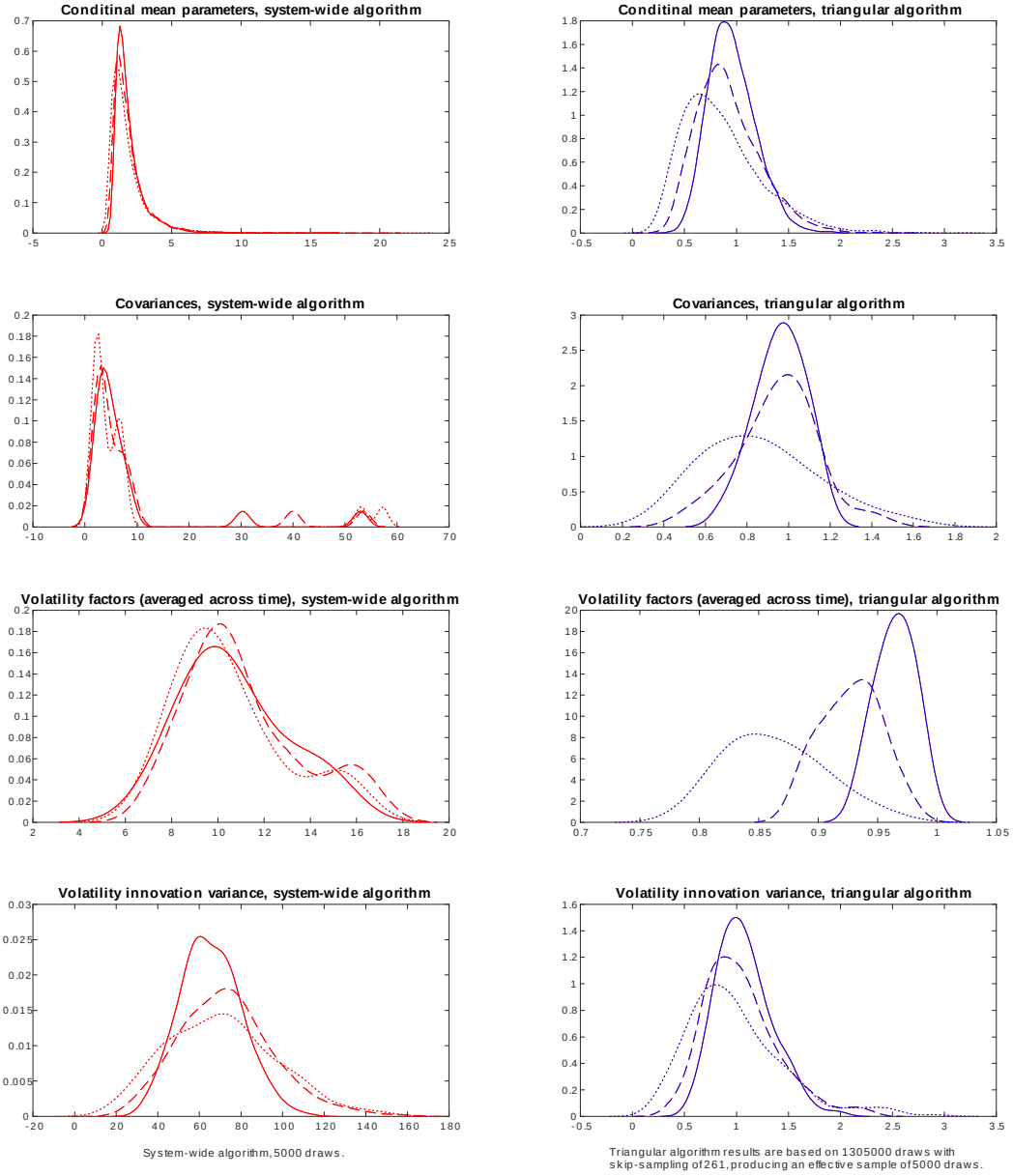

Figure 4: Comparison of Inefficiency Factors between the system wide and triangular algorithm. Kernel estimates. Solid, dashed, and dotted lines refer to 4, 8, and 15 percent tapering respectively. The densities in each sub-plot are computed across the parameters within a given set (from top to bottom: conditional mean coefficients, covariances, states, and covariances of the states). The graphs on the left refer to the system-wide algorithm, while the graphs on the right refer to the triangular algorithm.

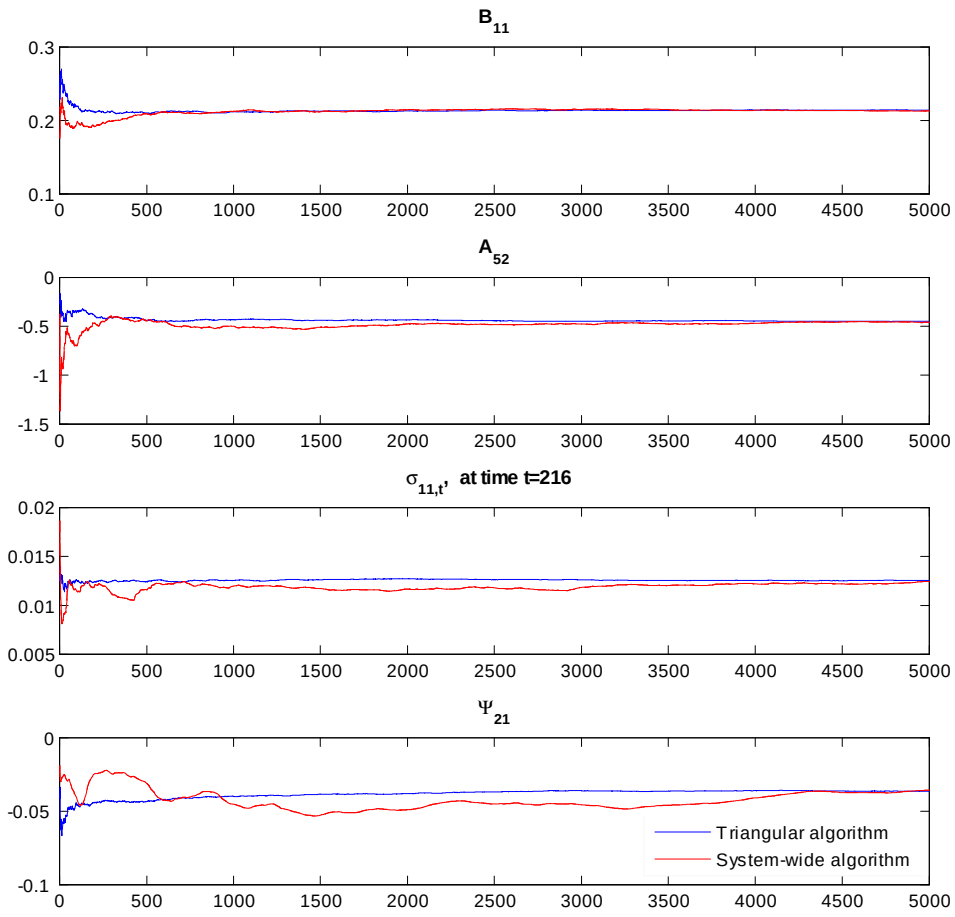

Figure 5: Recursive means of selected coefficients. Comparison between the system wide and triangular algorithm. The chains are initialised at the same value (set equal to the priors).

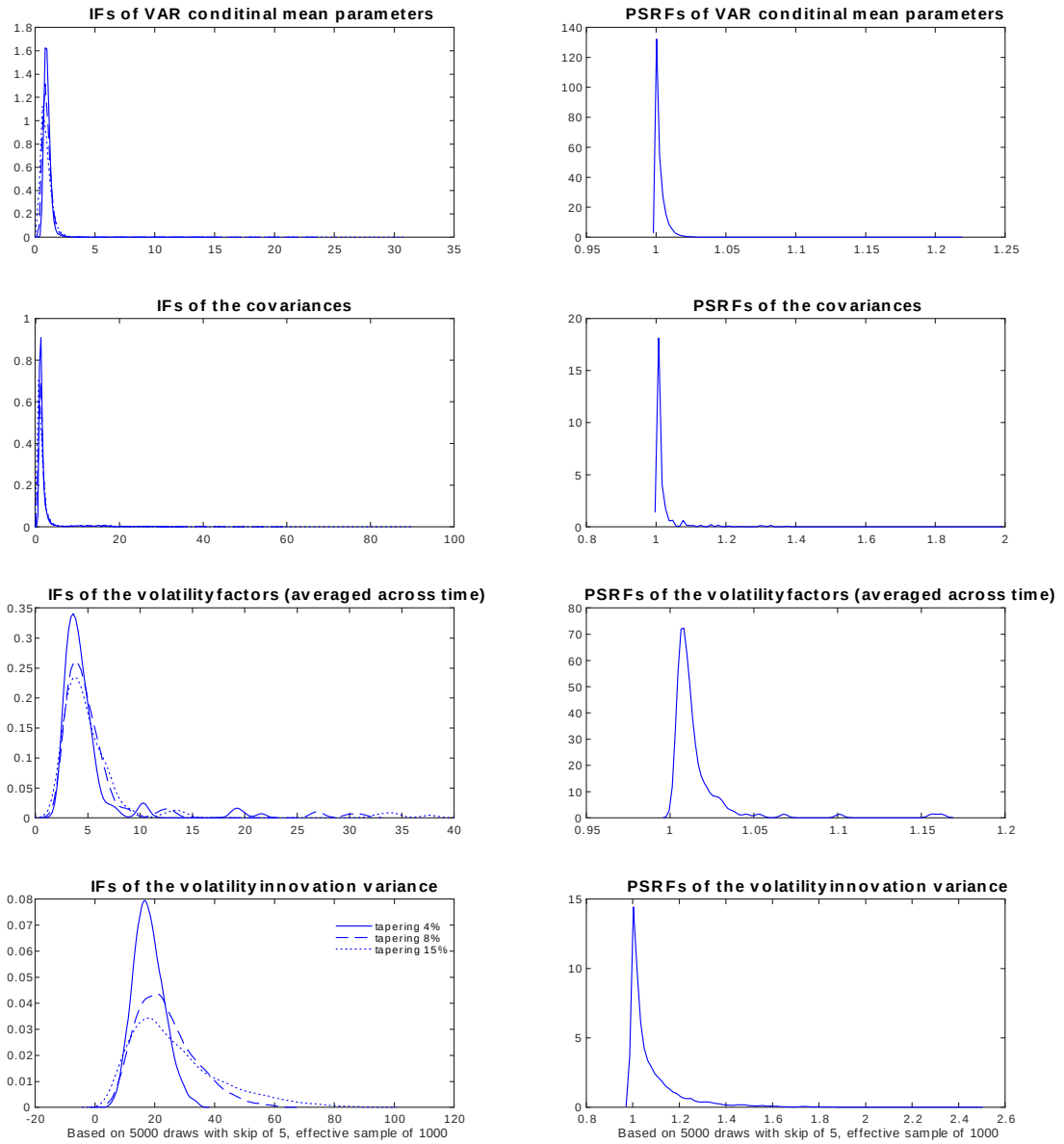

Figure 6: Convergence diagnostics. Inefficiency Factors (IF, panels on the left) and Potential Scale Reduction Factors (PSRF, panels on the right) of the 125-dimensional VAR coefficients and latent states. Kernel estimates. The densities in each subplot are computed across the parameters within a given set (from top to bottom: conditional mean coefficients, covariances, states, and covariances of the states).

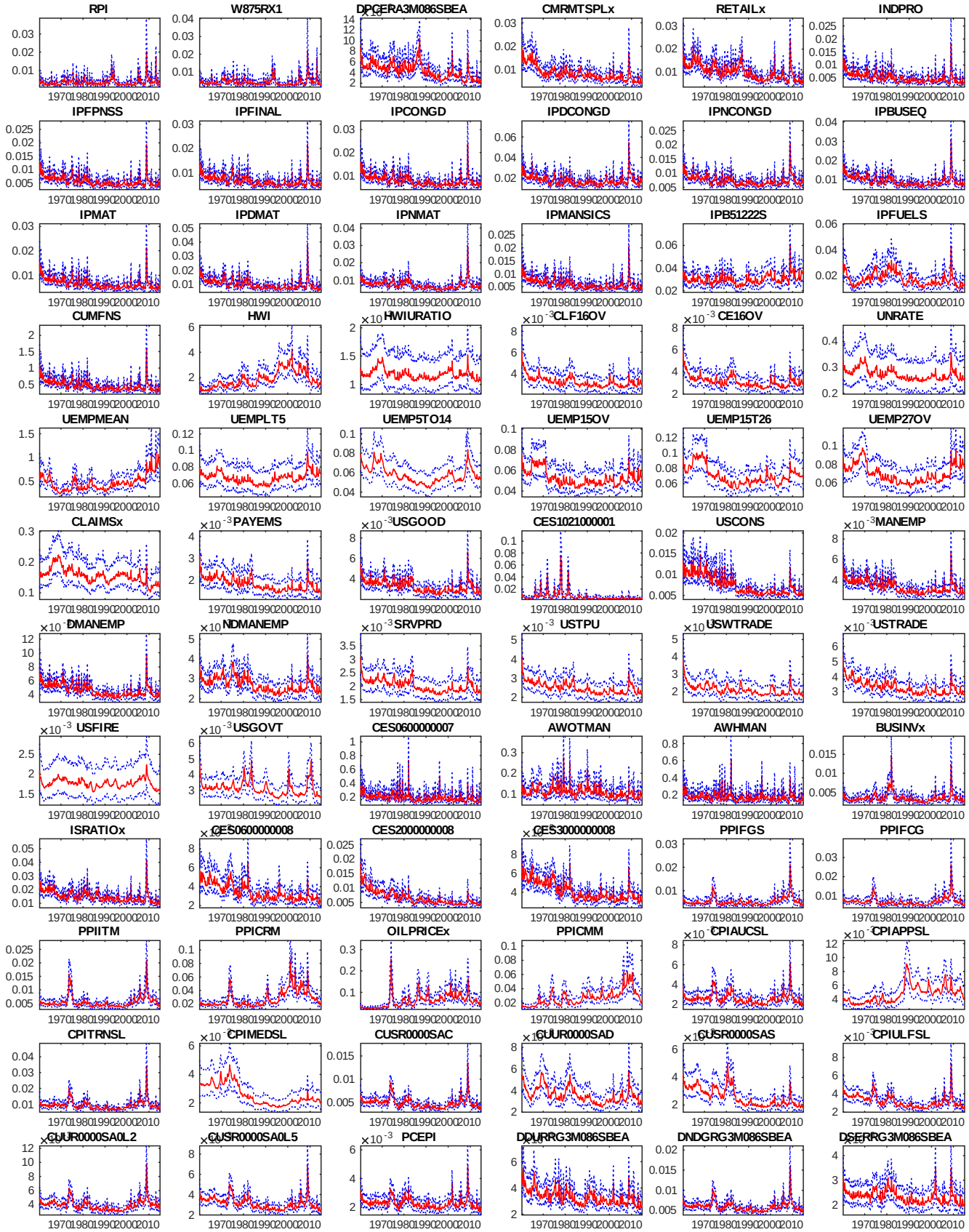

Figure 7: Posterior distribution of volatilities (diagonal elements of Σ_t), slow variables.

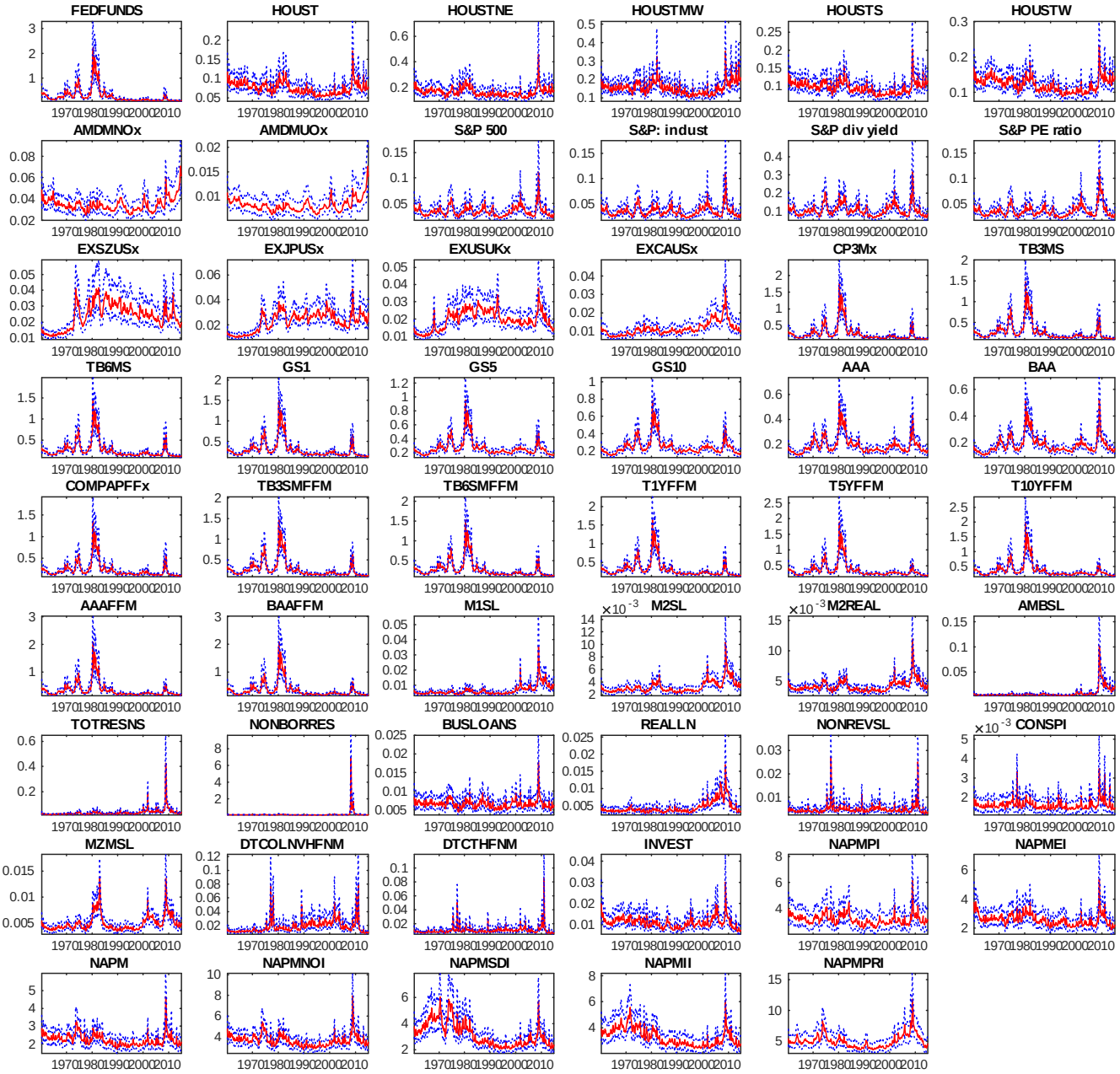

Figure 8: Posterior distribution of volatilities (diagonal elements of Σ_t), fast variables.

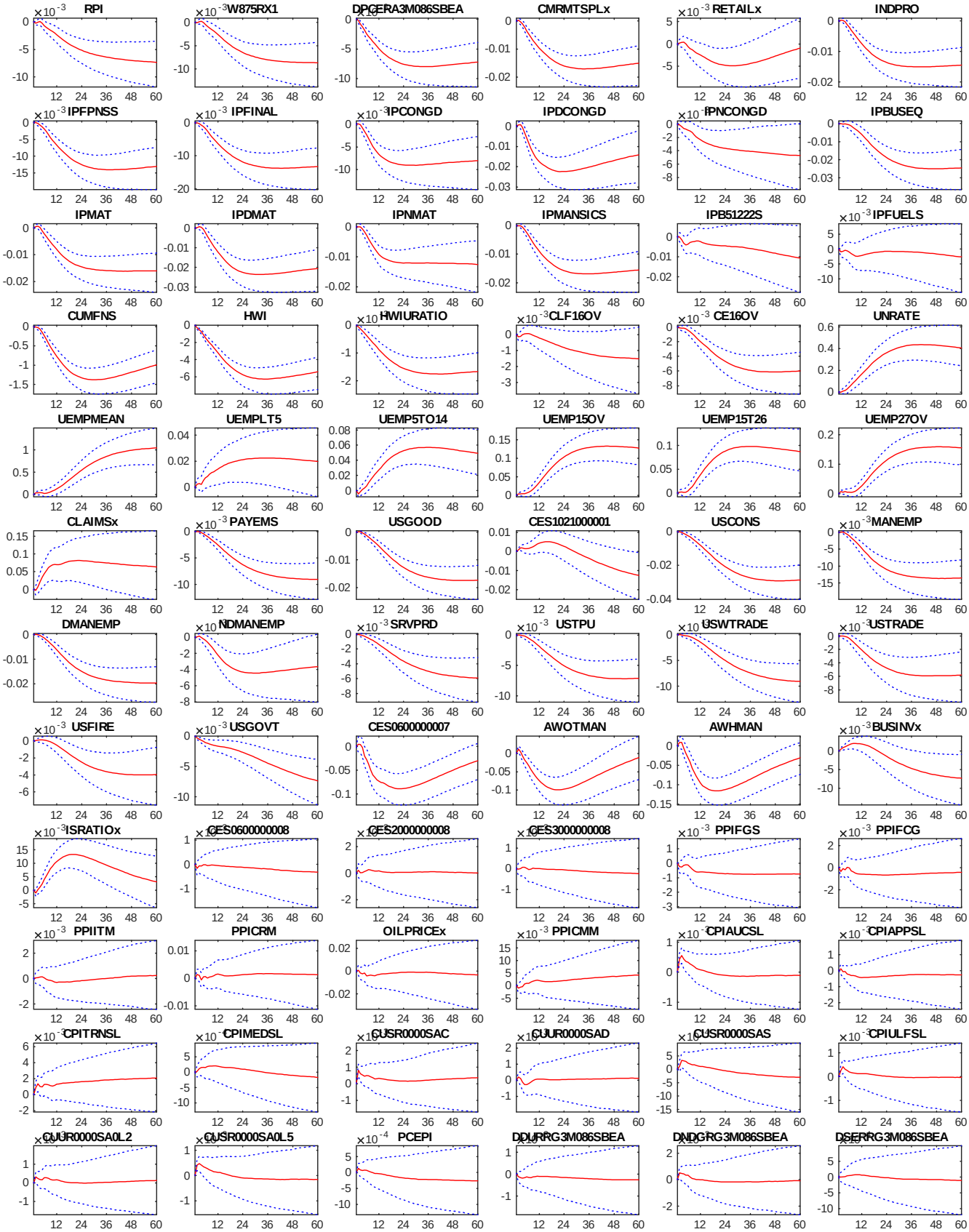

Figure 9: Impulse responses to a monetary policy shock: slow variables.

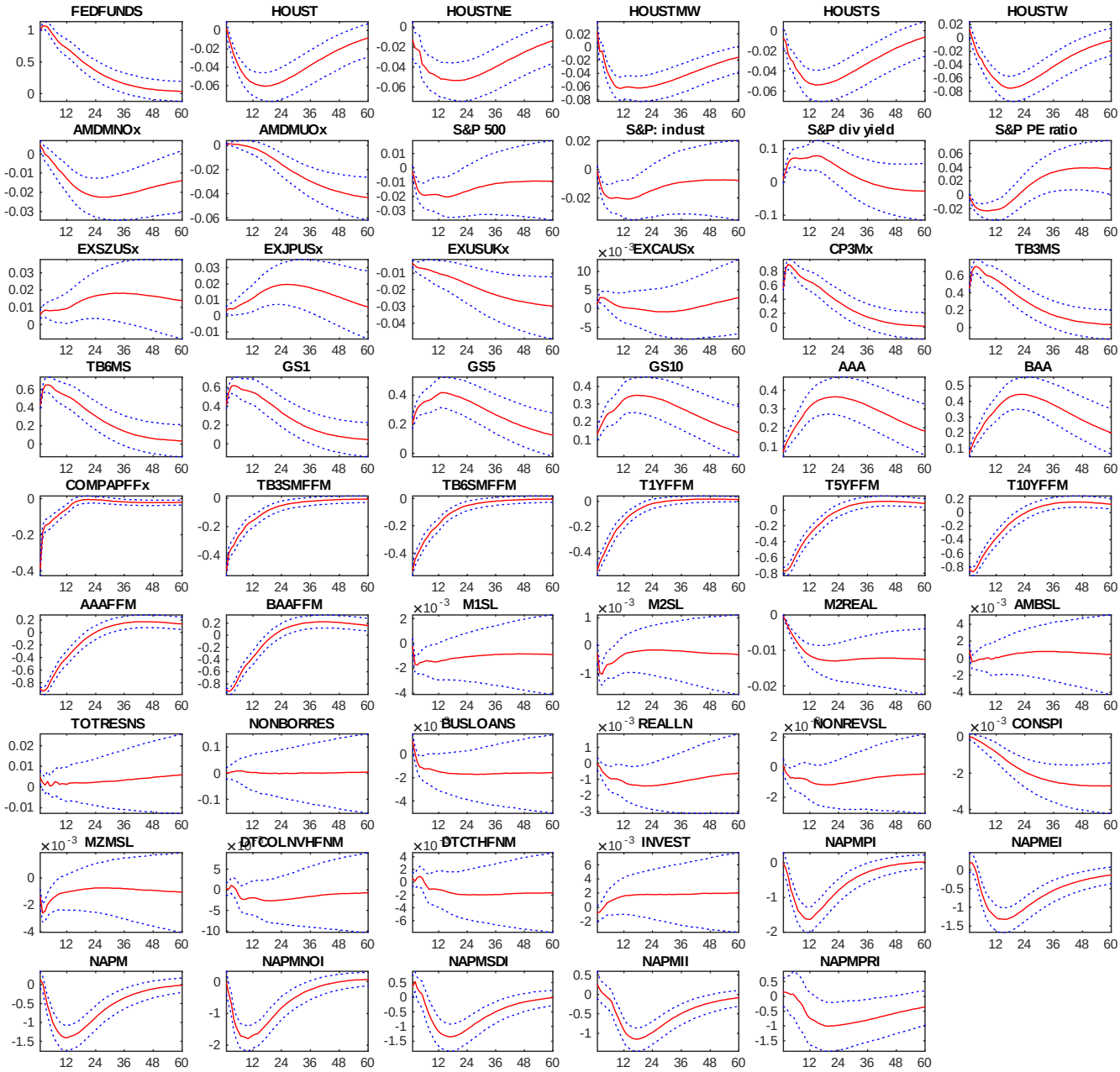

Figure 10: Impulse responses to a monetary policy shock: fast variables.

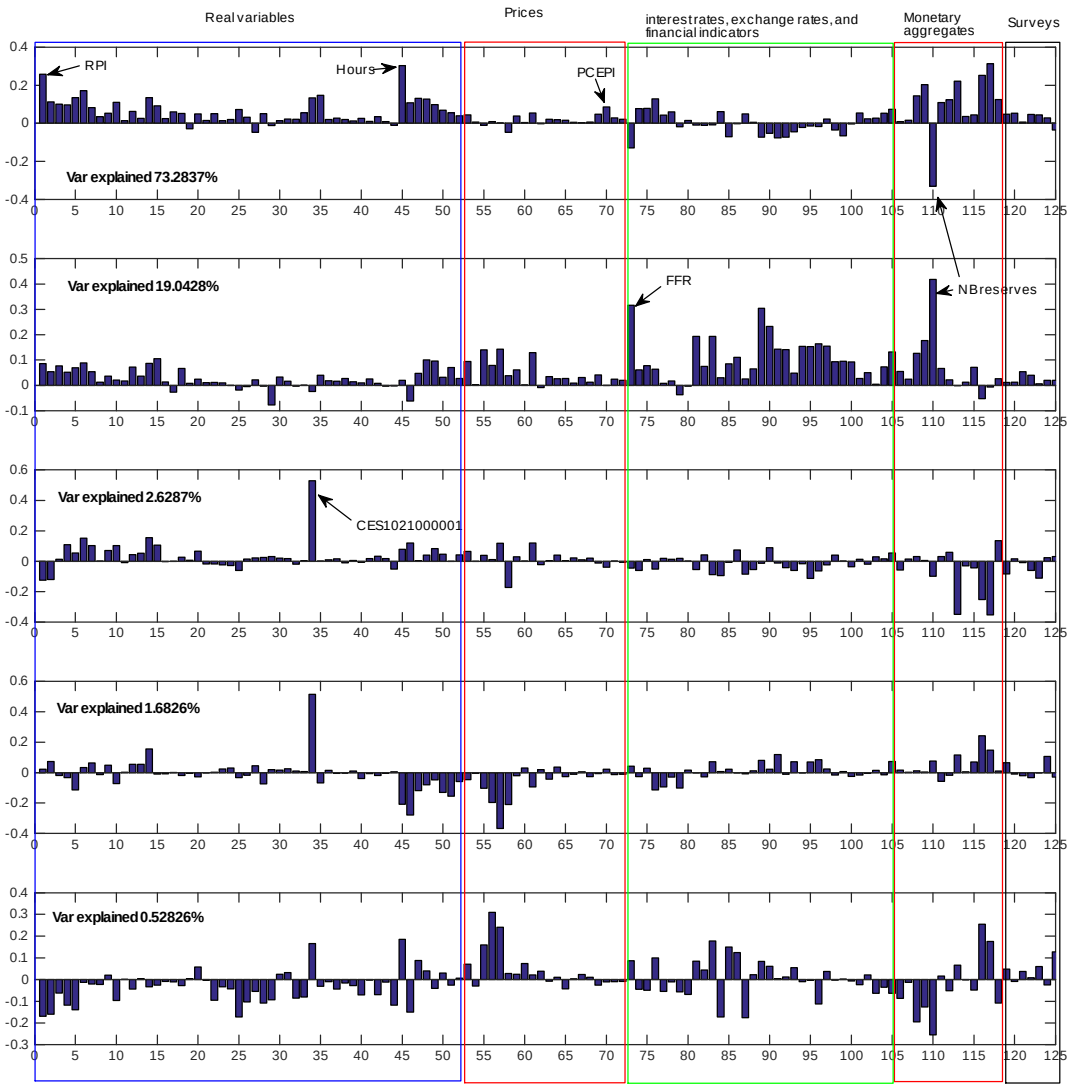

Figure 11: Principal components loadings of the variance-covariance of the volatilities (matrix Φ).

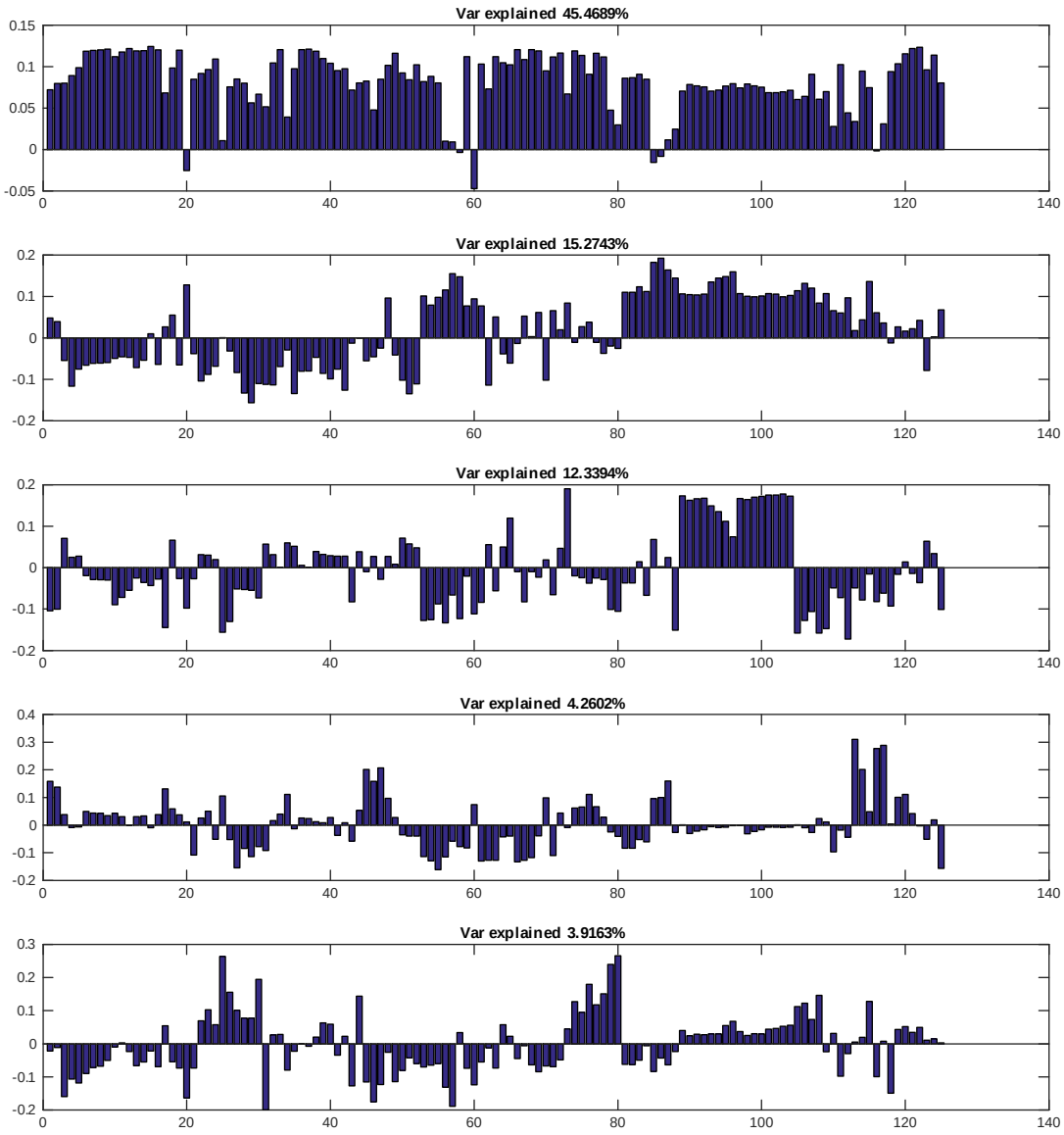

Figure 12: Principal components loadings of the volatility states.

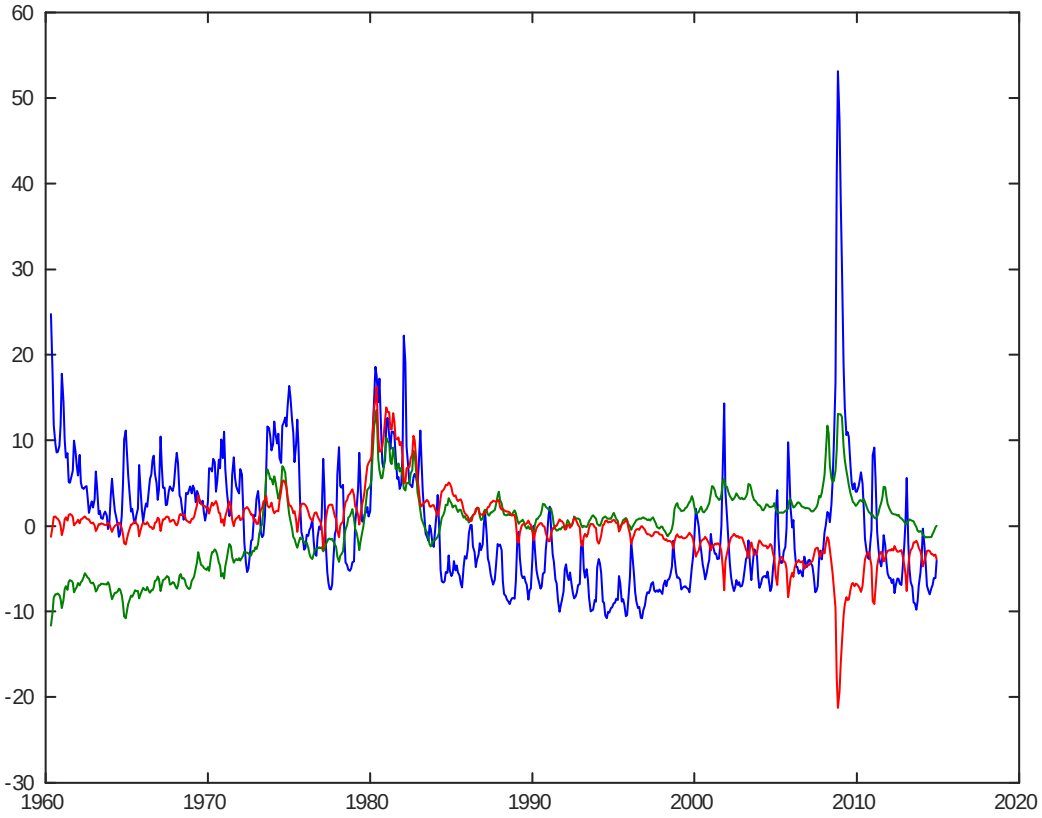

Figure 13: Common factors in volatilities. Factor 1 (blue) explains 45% of total variation, factor 2 (red) explains 15% of total variation, factor 3 (green) explains 12% of total variaiton.

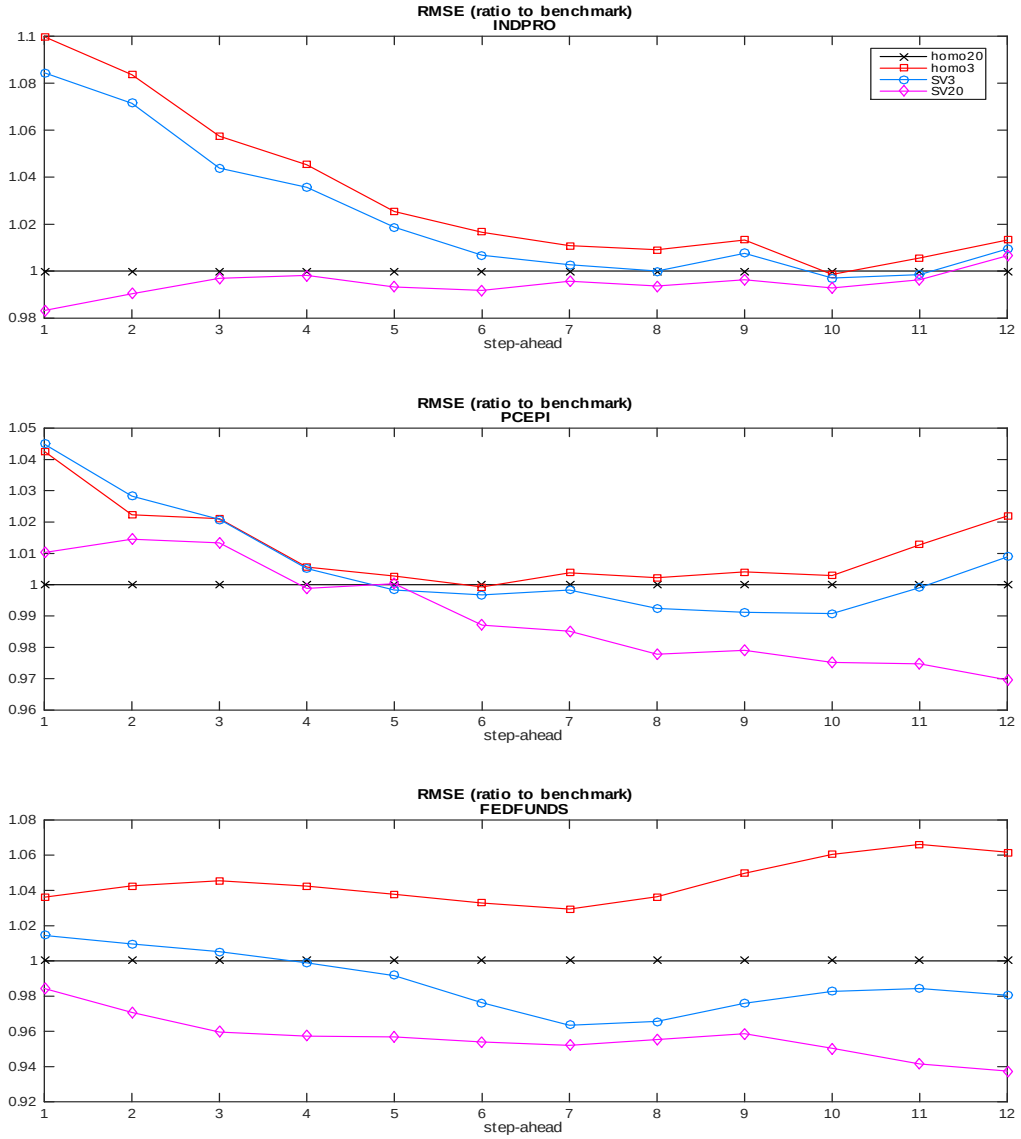

Figure 14: Point forecasts: relative RMSE of different models. Black line (benchmark, marker: crosses) is a homoschedastic VAR with 20 variables, red line (marker: squares) is a homoschedastic VAR with 3 variables, blue line (marker: circles) is heteroschedastic VAR with 3 variables, purple line (marker: diamonds) is heteroschedastic VAR with 20 variables.

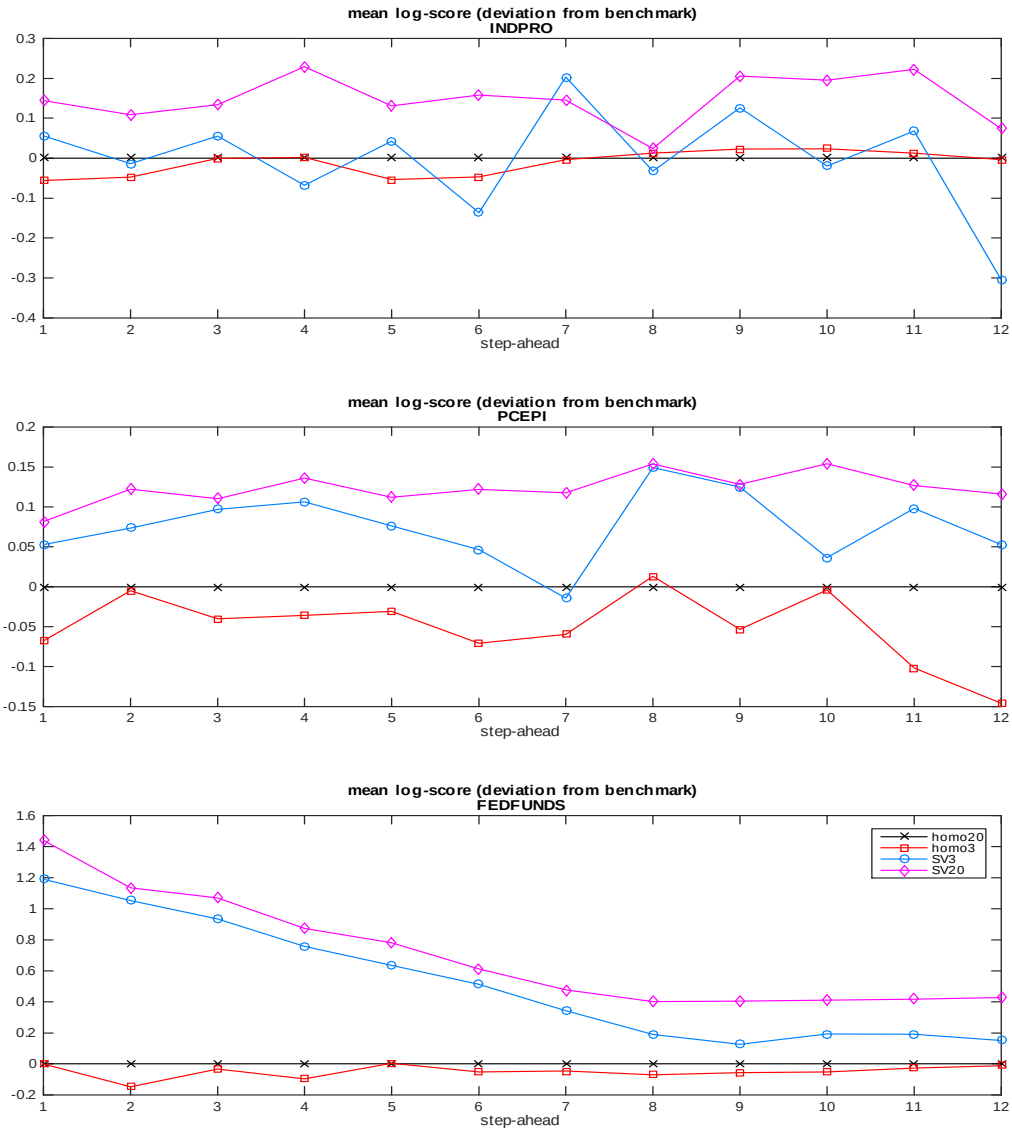

Figure 15: Point forecasts: Log-score gains of different models vs benchmark. Black line (benchmark, marker: crosses) is a homoschedastic VAR with 20 variables, red line (marker: squares) is a homoschedastic VAR with 3 variables, blue line (marker: circles) is heteroschedastic VAR with 3 variables, purple line (marker: diamonds) is heteroschedastic VAR with 20 variables.

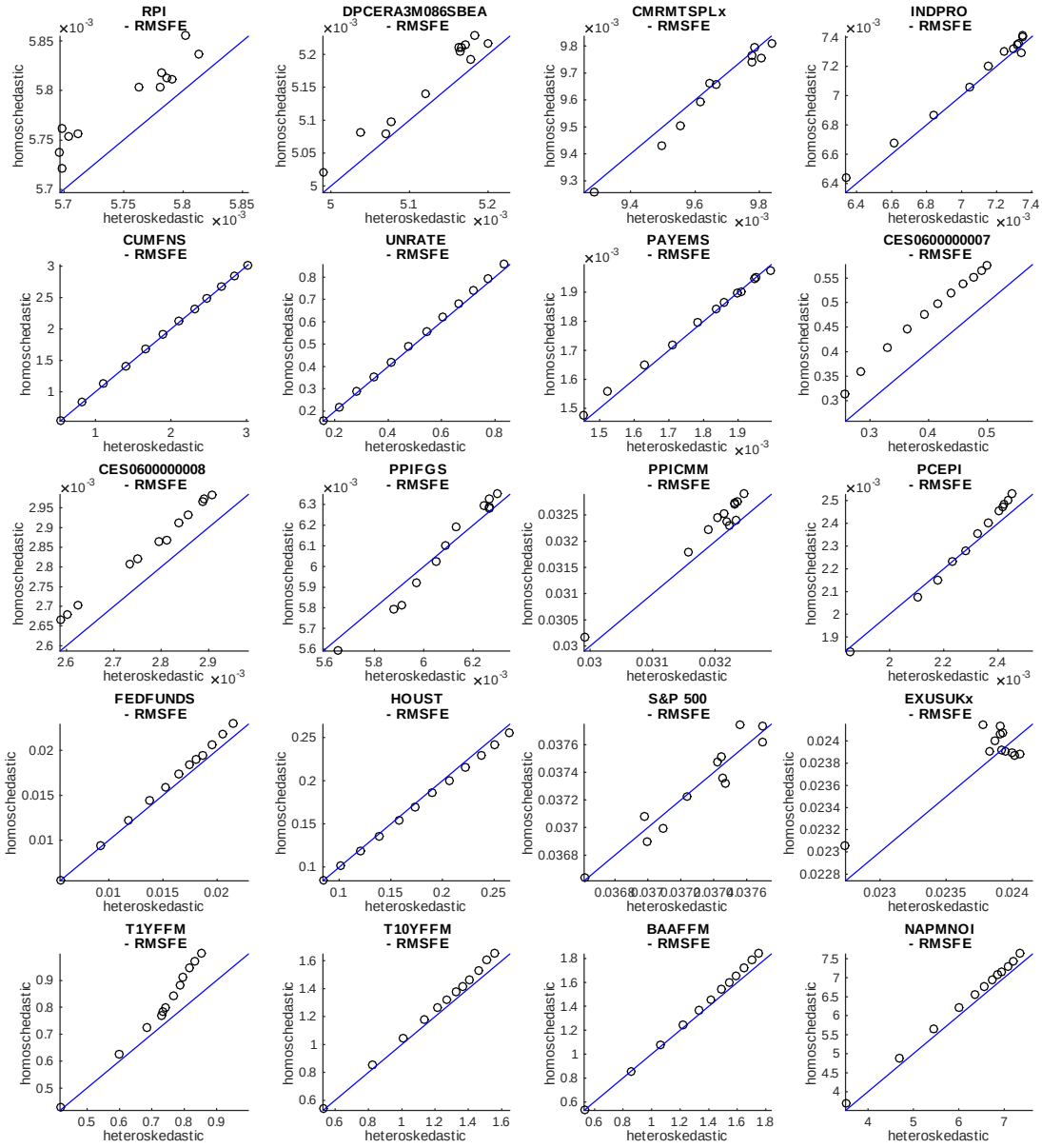

Figure 16: Comparison of point forecast accuracy. Each panel describes a different variable. The x axis reports the RMSFE obtained using the BVAR with stochastic volatility (heteroschedastic), the y axis reports the RMSFE obtained using the homoschedastic BVAR. Each point corresponds to a different forecast horizon from 1 to 12 step-ahead (in most cases, a higher RMSFE corresponds to a longer forecast horizon).

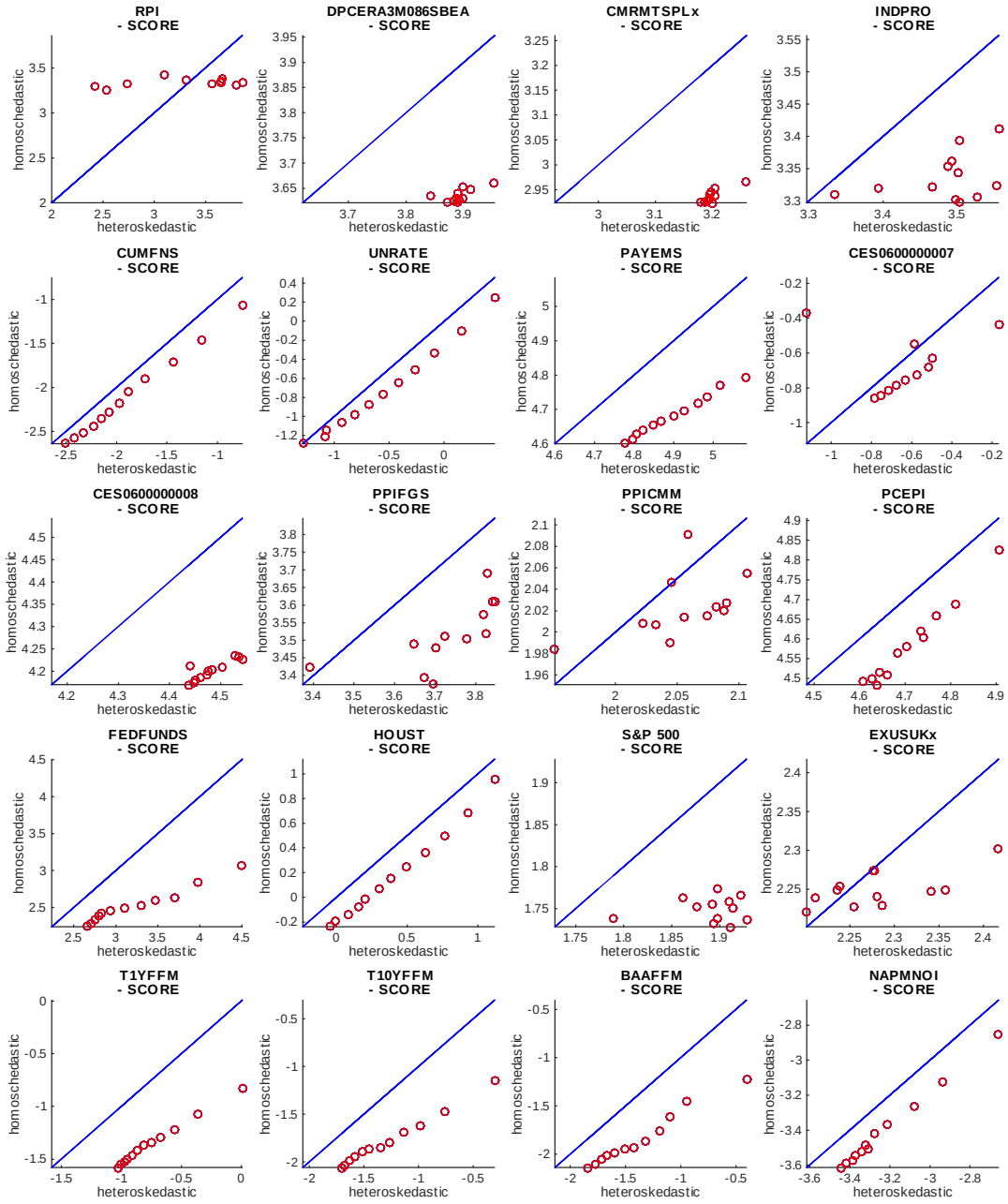

Figure 17: Comparison of density forecast accuracy. Each panel describes a different variable. The x axis reports the (log) density score obtained using the BVAR with stochastic volatility (heteroschedastic), the y axis reports the (log) density score obtained using the homoschedastic BVAR. Each point corresponds to a different forecast horizon from 1 to 12 step-ahead.