Large Vector Autoregressions with stochastic volatility and non-conjugate priors

Andrea Carriero [1] *Todd E. Clark* [2] *Massimiliano Marcellino* [3]

Norges Bank, 3 October 2017

[1] Queen Mary, University of London
[2] Federal Reserve Bank of Cleveland
[3] Bocconi University and CEPR

Introduction - two ingredients

- 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.** Banbura, Giannone, and Reichlin (2010), Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015) and Koop (2013)
- **Time variation in the volatilities.** Clark (2011), Clark and Ravazzolo (2015), Cogley and Sargent (2005), D'Agostino, Gambetti and Giannone (2013), and Primiceri (2005)
- There are no papers which jointly allow for **both** time variation **and** large datasets

Introduction - heteroskedasticity

- The reason lies in the structure of the likelihood function
- Homoskedastic VARs are SUR models with the same set of regressors in each equation $\longrightarrow$ Kronecker structure in the likelihood $\longrightarrow$ OLS equation by equation
- Equation-specific stochastic volatility breaks this symmetry because each equation is driven by a different volatility
- The system would need to be vectorised, and the conditional posterior involves manipulation of a matrix of dimension pN^2 (N=number of variables, p=number of lags)
- The computational complexity is therefore $N^{2^3} = N^6$

Introduction - asymmetric priors

- In a Bayesian framework, simmetry is not only needed in the likelihood, but also in the prior
- Kronecker structure in the likelihood+Kronecker structure in the prior= Kronecker structure in the posterior
- For example, the VAR estimated by Banbura, Giannone, and Reichlin (2010) is a VAR with 130 variables, but in order to make this estimation possible one needs to assume:
  - (i) Homoskedasticity of the disturbances
  - (ii) A specific structure for the prior
- Without either (i) or (ii) the system would need to be vectorised prior to estimation

The problem

- Consider the VAR of a N-dimensional vector y_t:

$$y_t = \Pi(L)y_{t-1} + v_t; \;\; v_t \sim iid \; N(0, \Sigma_t) \tag{1}$$

- Define $X_t = [1, y'_{t-1}, ..., y'_{t-p}]'$ and $\Pi = [\Pi_0|\Pi_1|...|\Pi_p]$
- In general we have the posterior $\text{vec}(\Pi)|\Sigma, y \sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi)$ with posterior precision:

$$\bar{\Omega}_\Pi^{-1} = \underbrace{\underline{\Omega}_\Pi^{-1}}_{\text{Prior}} + \underbrace{\sum_{t=1}^{T} (\Sigma_t^{-1} \otimes X_t X_t')}_{\text{Likelihood}} \tag{2}$$

- The precision matrix $\bar{\Omega}_\Pi^{-1}$ is of size $N(Np+1)$. Its manipulation requires $(pN^2)^3 = O(N^6)$ elementary operations
- For N very large modern computers (laptops/desktops) can't even store such a matrix in RAM (e.g. $N = 125$ needs 330 GB of RAM).

The usual solution

- In general we have the posterior $\text{vec}(\Pi)|\Sigma, y \sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi)$ with

$$\bar{\Omega}_\Pi^{-1} = \underbrace{\underline{\Omega}_\Pi^{-1}}_{\text{Prior}} + \sum_{t=1}^{T} \underbrace{(\Sigma_t^{-1} \otimes X_t X_t')}_{\text{Likelihood}} \tag{2}$$

- Now assume that

  (i) $\Sigma_t = \Sigma$ (homoskedasticity)

  (ii) $\underline{\Omega}_\Pi = \Sigma \otimes \Omega_0$ (conjugate prior)

$$\bar{\Omega}_\Pi^{-1} = \underbrace{\underline{\Omega}_\Pi^{-1}}_{\Sigma^{-1}\otimes\Omega_0^{-1}} + \sum_{t=1}^{T} (\underbrace{\Sigma_t^{-1}}_{\Sigma^{-1}} \otimes X_t X_t') = \Sigma^{-1} \otimes \left(\Omega_0^{-1} + \sum_{t=1}^{T} X_t X_t' \right), \tag{3}$$

  and the two terms can be manipulated separately, reducing complexity by $O(N^3)$

- Classical homoskedastic VARs can be estimated equation by equation.

Problems with the the usual solution

The Natural-conjugate homoskedastic approach allows to use large datasets, but it has important limitations:

- It imposes homoskedasticity, against the overwhelming evidence in macroeconomic and financial data
- The prior structure $\Sigma \otimes \Omega_0$ is restrictive (Rothemberg (1963), Sims and Zha (1998))
  - It prevents any asymmetry in the prior across equations, because the coefficients of each equation feature the same prior variance Ω_0 (up to a scale factor given by the elements of Σ).
  - It has the unappealing consequence that prior beliefs must be correlated across equations, with a correlation structure proportional to that of the shocks (as described by Σ).

A new algorithm

- In this paper we propose a new algorithm that makes possible to use:
  - A heteroskedastic model
  - The more general and less restrictive independent Normal - Inverse Wishart (and Normal-diffuse) prior
- Our procedure is based on a simple **factorization of the likelihood**, which allows to draw the VAR coefficients equation by equation
- This reduces the computational complexity from N^6 to N^4.
- Our new algorithm is very simple and can be easily inserted in any pre-existing algorithm for estimation of BVAR models.

The Model

- Consider the following VAR model for a N-dimensional y_t 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 Λ_t is a diagonal matrix with generic j-th element $h_{j,t}$ and A^{-1} is a lower triangular matrix with ones on its main diagonal.

- The bottleneck is drawing $\text{vec}(\Pi)|A, \Lambda_T, y_T \sim N(\text{vec}(\bar{\mu}_\Pi), \overline{\Omega}_\Pi)$; To obtain a draw one needs to i) invert

$$\bar{\Omega}_\Pi^{-1} = \underbrace{\underline{\Omega}_\Pi^{-1}}_{\text{Prior}} + \underbrace{\sum_{t=1}^{T} (\Sigma_t^{-1} \otimes X_t X_t')}_{\text{Likelihood}} \quad (3)$$

  ii) compute its Cholesky factor and iii) multiply the Cholesky factor by a random vector

- Each of the above operations is of complexity N^6

An algorithm for large VARs

Consider again the decomposition $v_t = A^{-1}\Lambda_t^{0.5}\epsilon_t$:

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

where $a^*_{j,i}$ denotes the generic element of the matrix A^{-1} which is available under knowledge of A.

An algorithm for large VARs

The VAR can be written as:

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

with the generic equation for variable j:

$$
\underbrace{y_{j,t} - (a_{j,1}^{*} h_{1,t}^{0.5}\epsilon_{1,t} + \dots + a_{j,,j-1}^{*} h_{j-1,t}^{0.5}\epsilon_{j-1,t})}_{y_{j,t}^{*}} = \pi_j^{(0)} + \sum_{i=1}^{N}\sum_{l=1}^{p} \pi_{j,l}^{(i)} y_{i,t-l} + h_{j,t}\epsilon_{j,t}. \quad (4)
$$

When drawing the coefficients of equation j the term $y_{j,t}^{*}$ is known, since it is given by the difference between the dependent variable of that equation and the realized residuals of all the previous $j-1$ equations. Hence (4) is a standard generalized linear regression model with i.i.d. Gaussian disturbances.

An algorithm for large VARs

The full conditional posterior distribution of the conditional mean coefficients can be factorized as:

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

and one can draw the coefficients in Π in separate blocks:

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

with

$$\begin{aligned} \bar{\mu}_{\Pi^{\{j|1:j-1\}}} &= \overline{\Omega}_{\Pi^{\{j|1:j-1\}}}\left\{\sum_{t=1}^{T} X_{j,t}h_{j,t}^{-1}y_{j,t}^{*\prime}+\underline{\Omega}_{\Pi^{\{j|1:j-1\}}}^{-1}\underline{\mu}_{\Pi^{\{j|1:j-1\}}}\right\} \\ \overline{\Omega}_{\Pi^{\{j|1:j-1\}}}^{-1} &= \underline{\Omega}_{\Pi^{\{j|1:j-1\}}}^{-1}+\sum_{t=1}^{T} X_{j,t}h_{j,t}^{-1}X_{j,t}^{\prime}, \end{aligned}$$

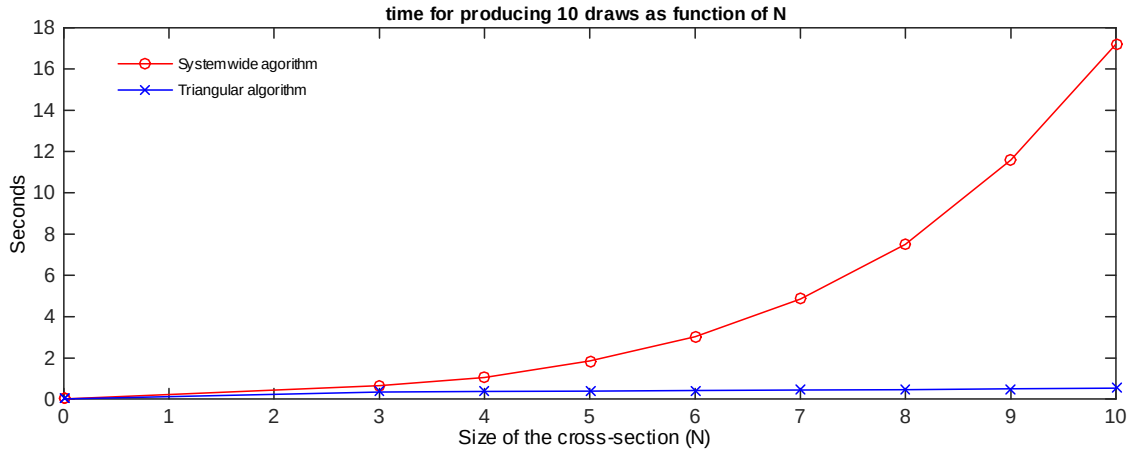
where $\underline{\mu}_{\Pi^{\{j|1:j-1\}}}$ and $\underline{\Omega}_{\Pi^{\{j|1:j-1\}}}$ are moments of $\Pi^{\{j\}}|\Pi^{\{1:j-1\}} \sim N(\underline{\mu}_{\Pi^{\{j|1:j-1\}}},\underline{\Omega}_{\Pi^{\{j|1:j-1\}}})$

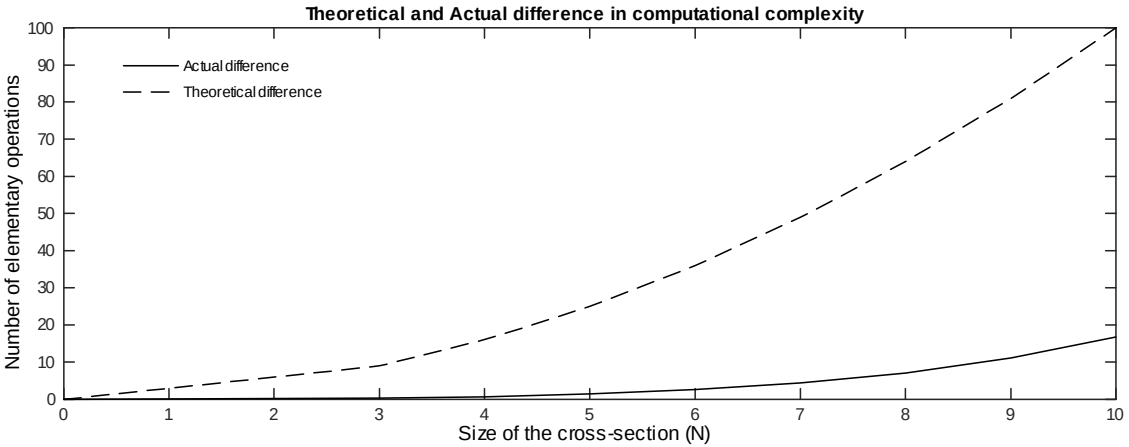
An algorithm for large VARs

- The conditional posterior of Π obtained is the **same** as the one from the system-wide algorithm
  - The algorithm will produce draws **numerically identical** to those of the system-wide sampler
  - This is true regardelss of the **ordering**, which **is irrelevant** to the conditional posterior of Π
- The total computational complexity of this estimation algorithm is $O(N^4)$, with a gain of N^2.
  - Uses equations with at most $Np+1$ regressors, and the correlation across equations typical of SUR models is implicitly accounted for by the factorization
  - The dimension of the posterior variance matrix $\overline{\Omega}_{\Pi\{j\}}^{-1}$ is $(Np+1)$, which means that its manipulation only involves operations of order $O(N^3)$.

Computational complexity and speed of simulation

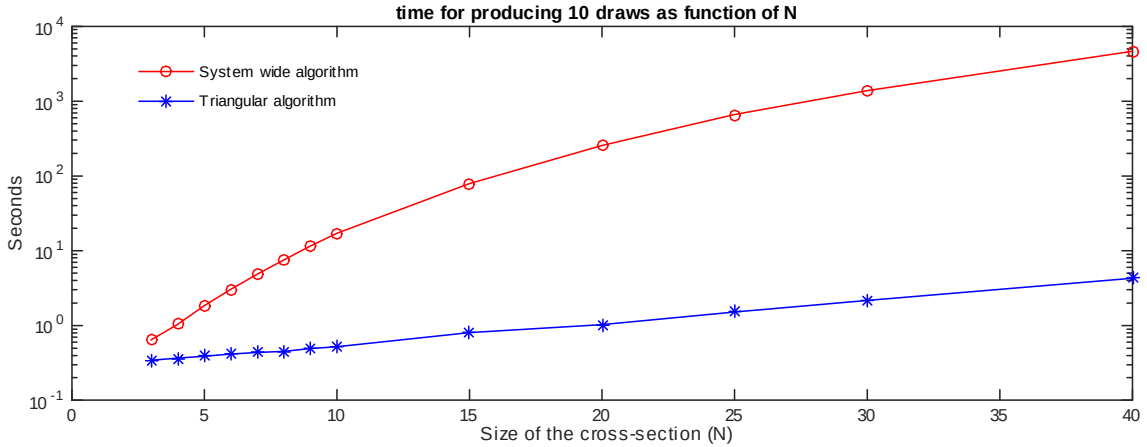
time for producing 10 draws as a function of N

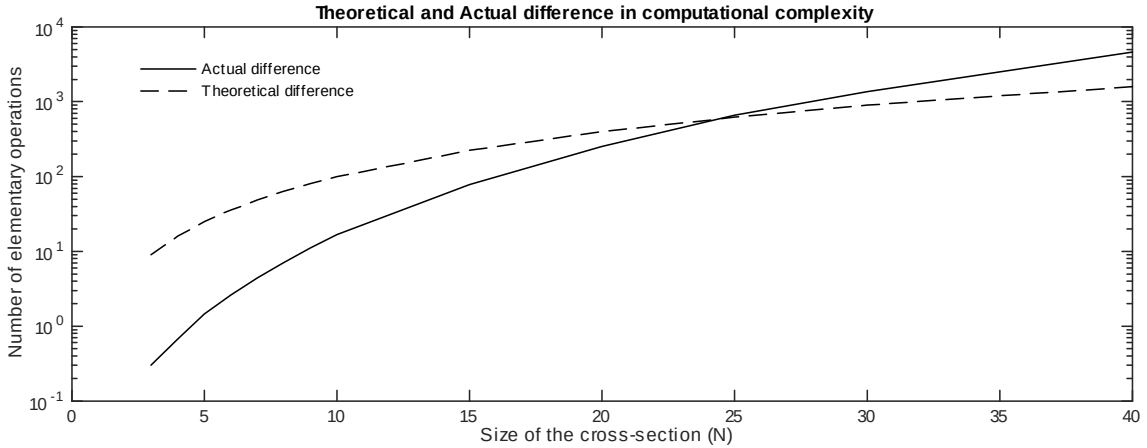




Computational complexity and speed of simulation

time for producing 10 draws as a function of N - log scale



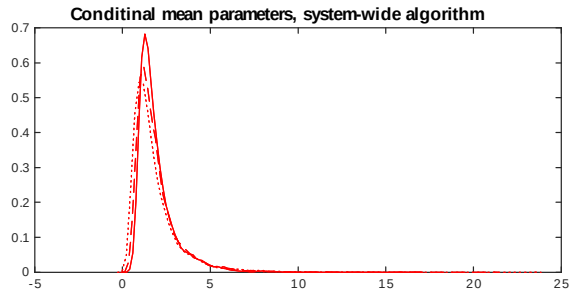


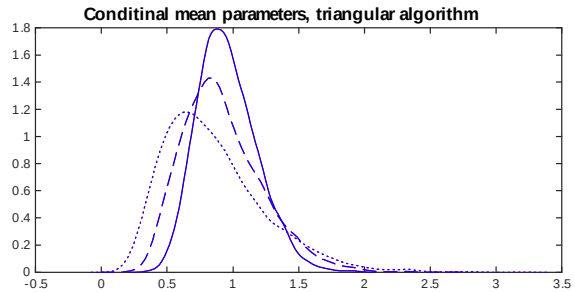
Convergence and mixing

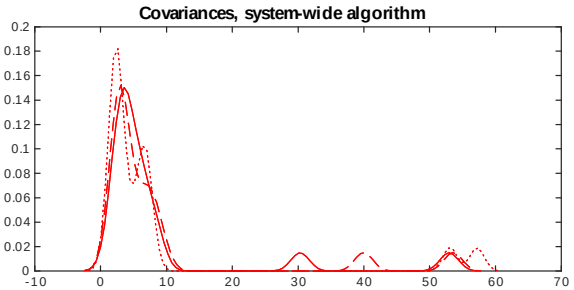
- Regardless of the power of the computers used to perform the simulation the triangular algorithm will always produce many more draws than the traditional system-wide algorithm **in a given unit of time**.
- This has important consequences in terms of producing draws with good mixing and convergence properties.
- The triangular algorithm can produce draws **many times closer to i.i.d. sampling** in the same amount of time.
- These computational and storage gains increase quadratically with the system size

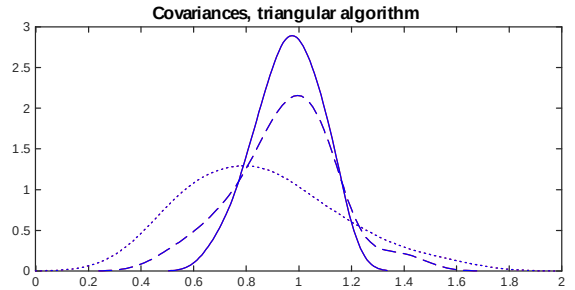
Convergence and mixing

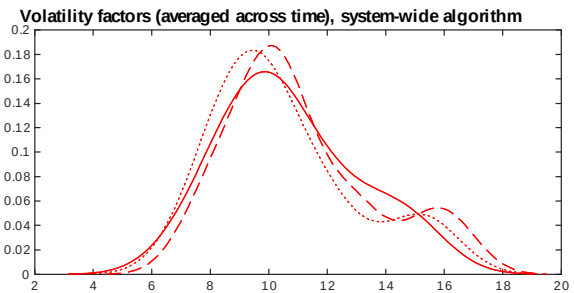
Inefficiency factors= distance from i.i.d sampling: ideally should be around 1.

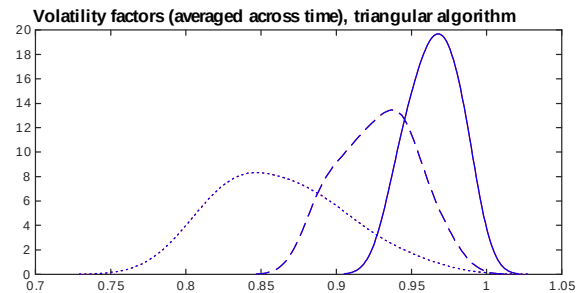












Empirical applications

- As an illustration we estimate a VAR with stochastic volatilities, using 13 lags and a cross-section of 125 variables from FRED-MD
- For a model of this size the system-wide algorithm would have a covariance matrix of the coefficients of dimension 203250, which would require about 330 GB of RAM ($203250^2 \times 8/10^9$).
- Our estimation algorithm can produce 5000 draws in just above 7 hours on a 3.5 GHz Intel Core i7.
- We find that:
  - The variance of the shocks was clearly unstable over time
  - There is a factor structure in the volatilities
  - The combined use of both time variation in volatilities and a large data-set improves point and density forecasts, more that what these two ingredients do if used separately.

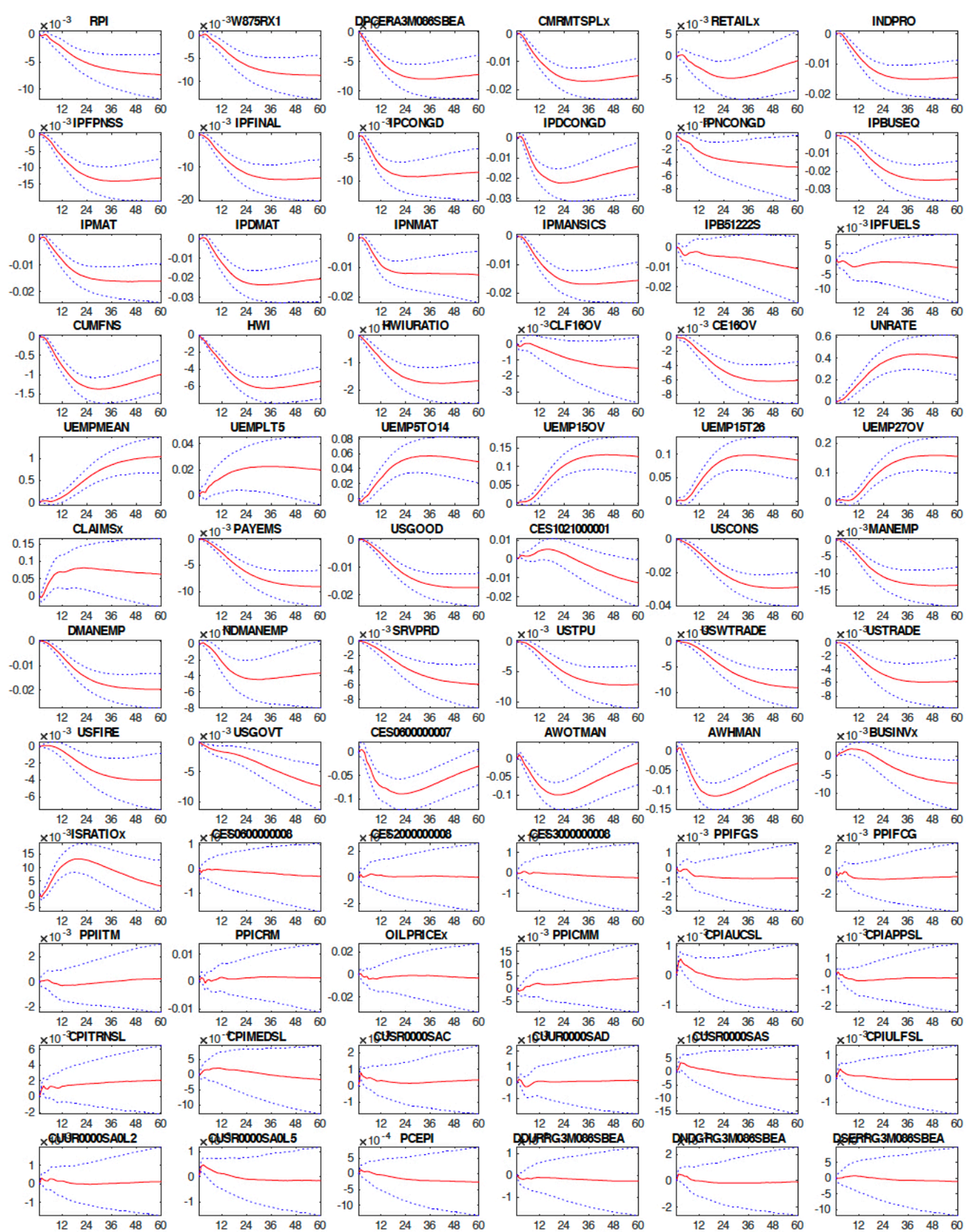


Figure 9: Impulse responses to a monetary policy shock: slow variables.

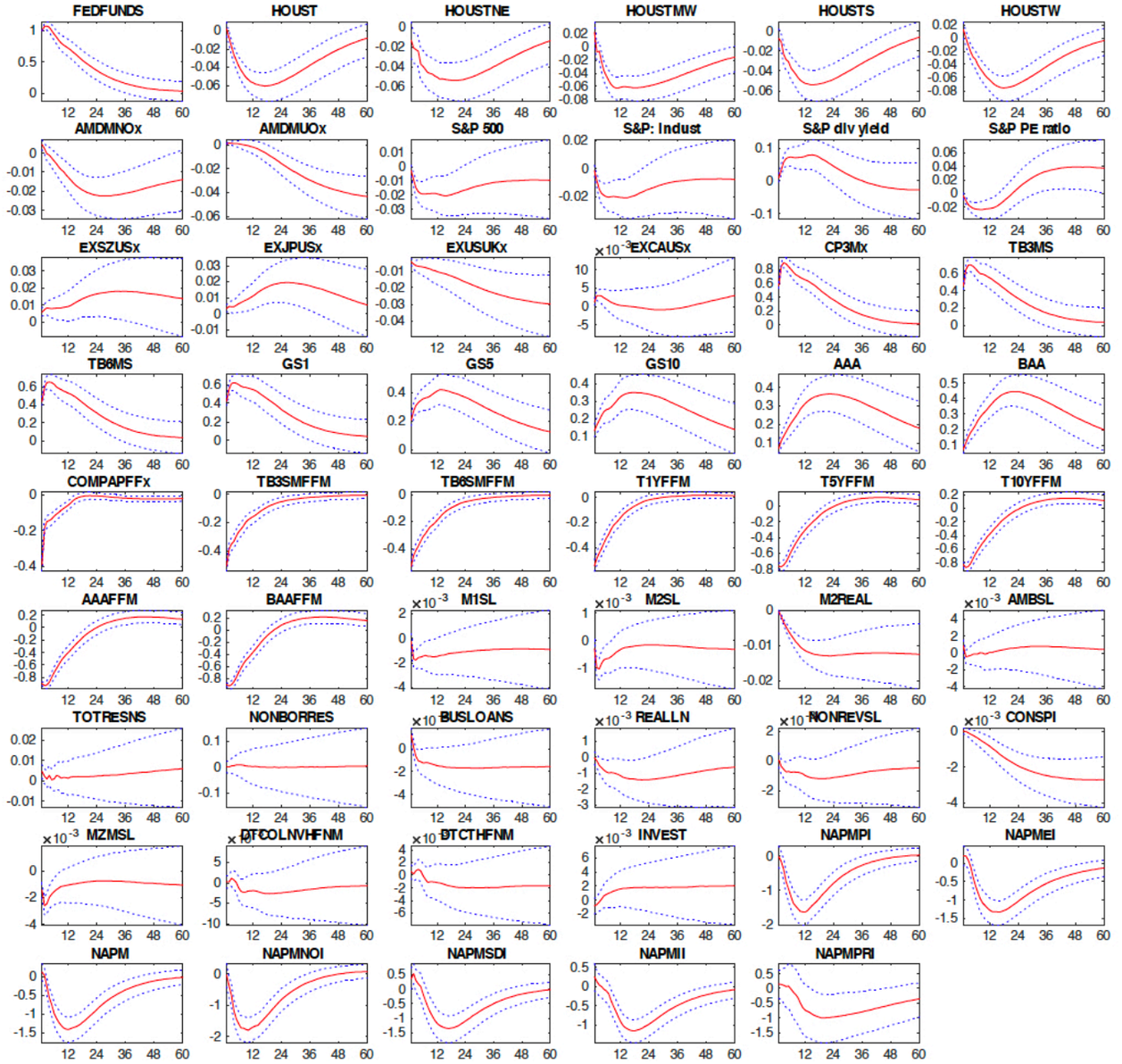


Figure 10: Impulse responses to a monetary policy shock: fast variables.

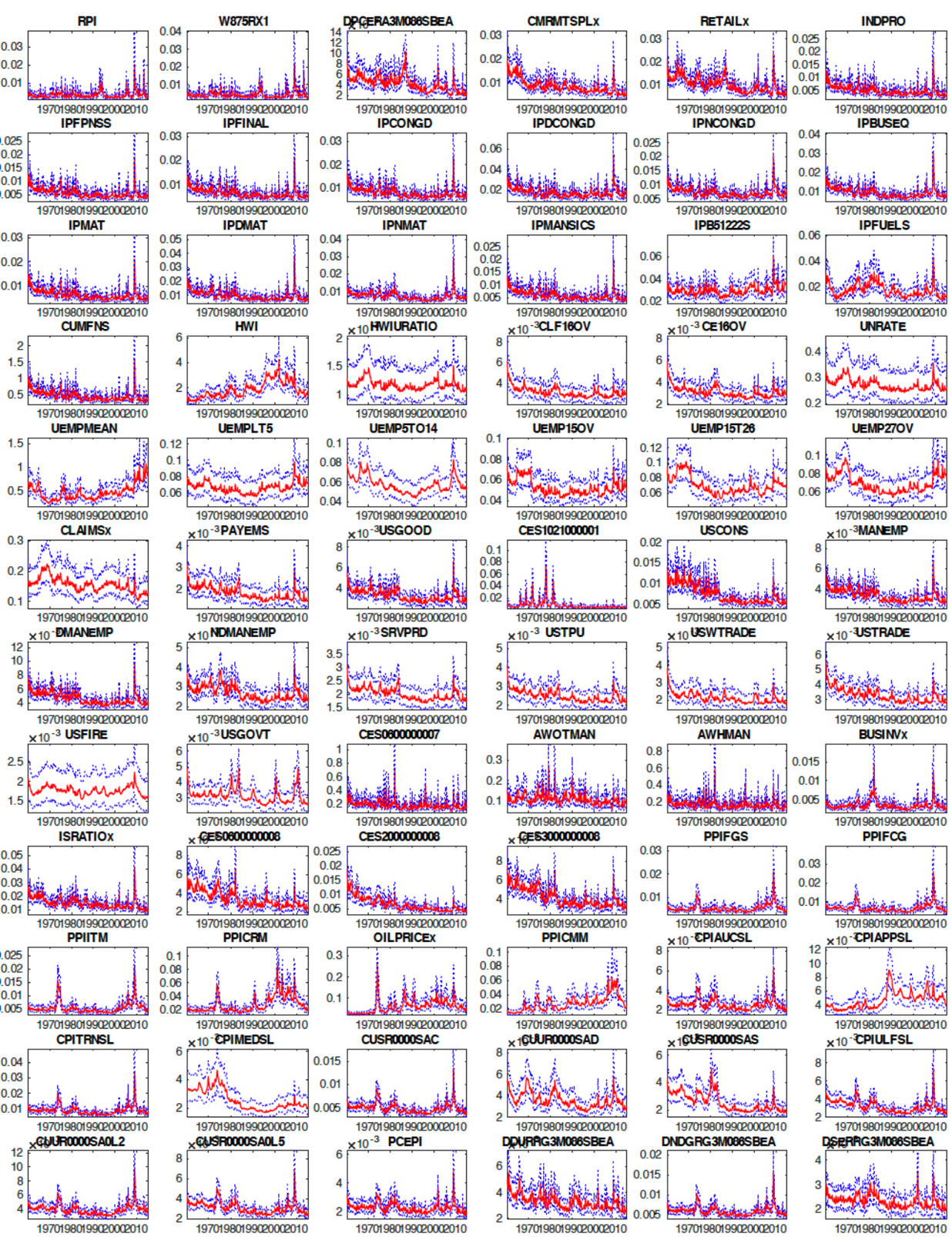


Figure 7: Posterior distribution of volatilities (diagonal elements of Σ_t), slow variables.

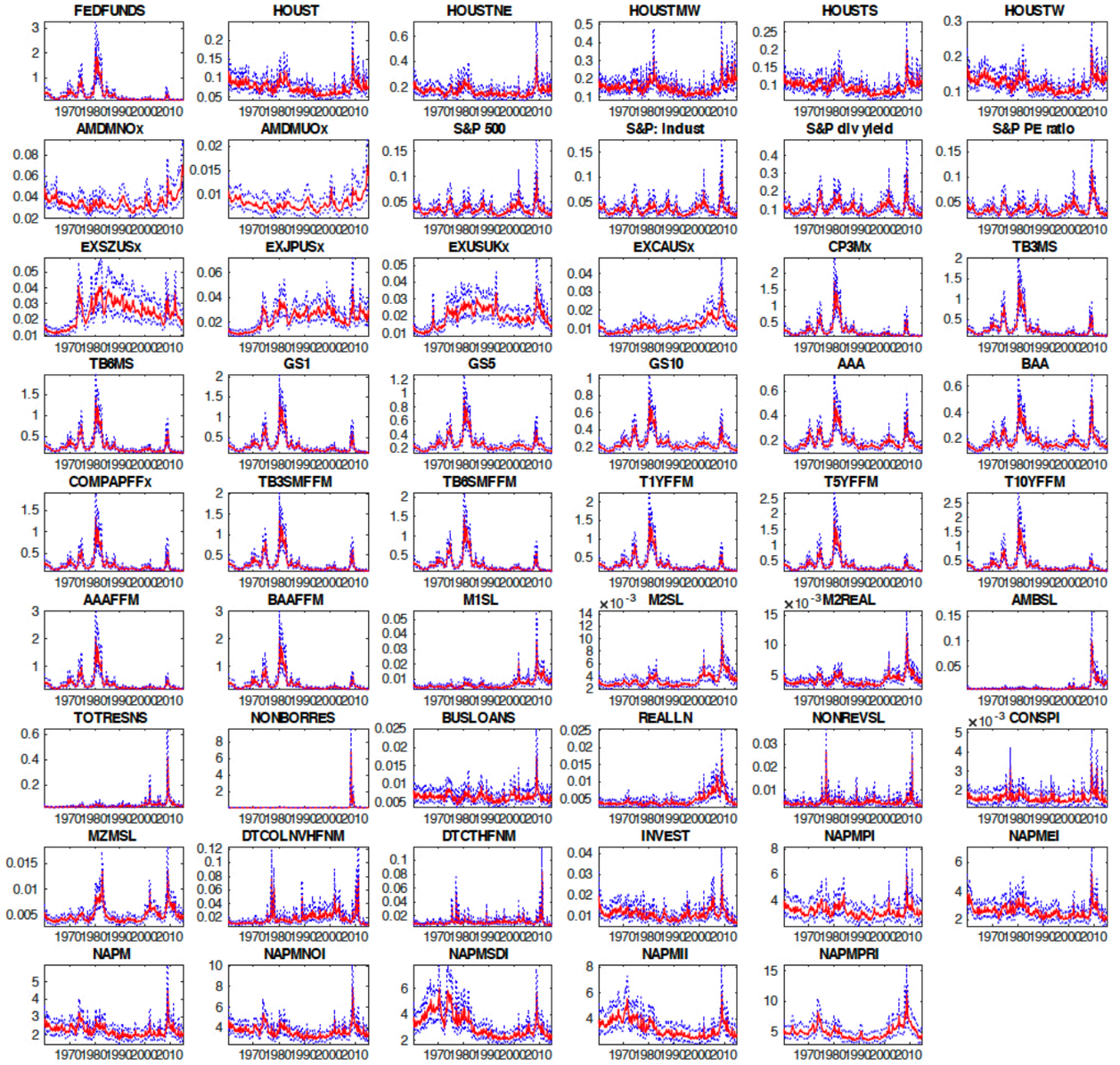
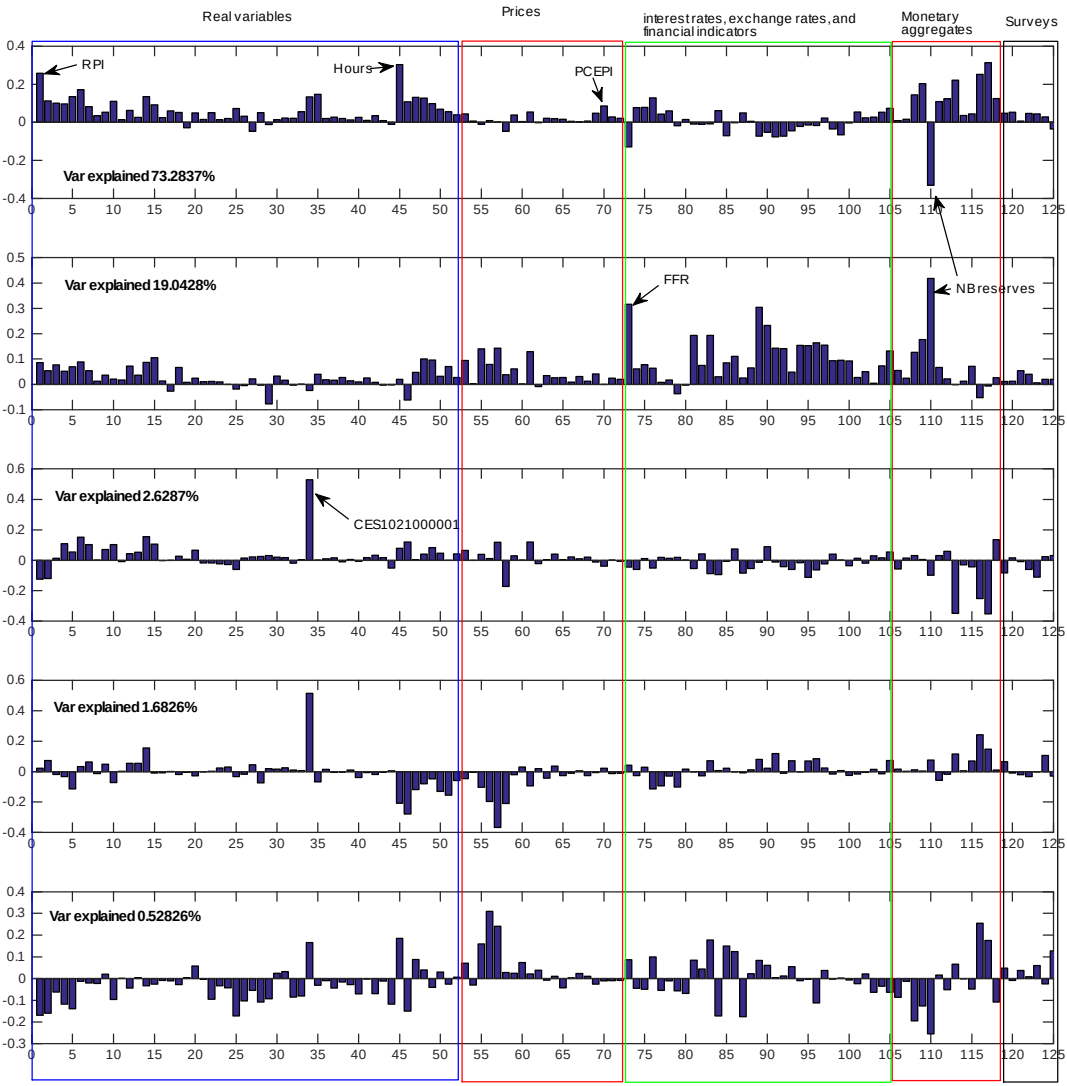
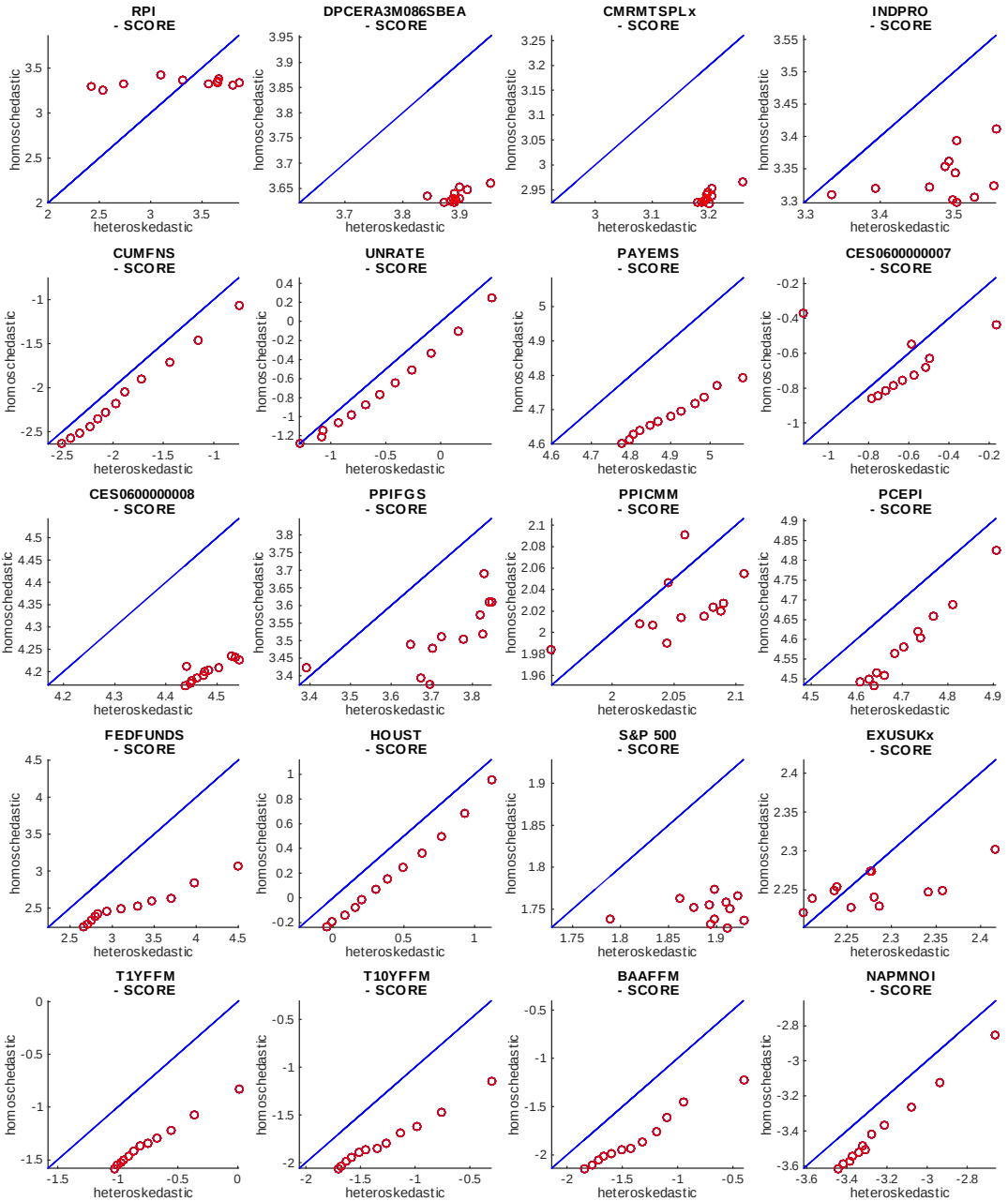


Figure 8: Posterior distribution of volatilities (diagonal elements of Σ_t), fast variables.

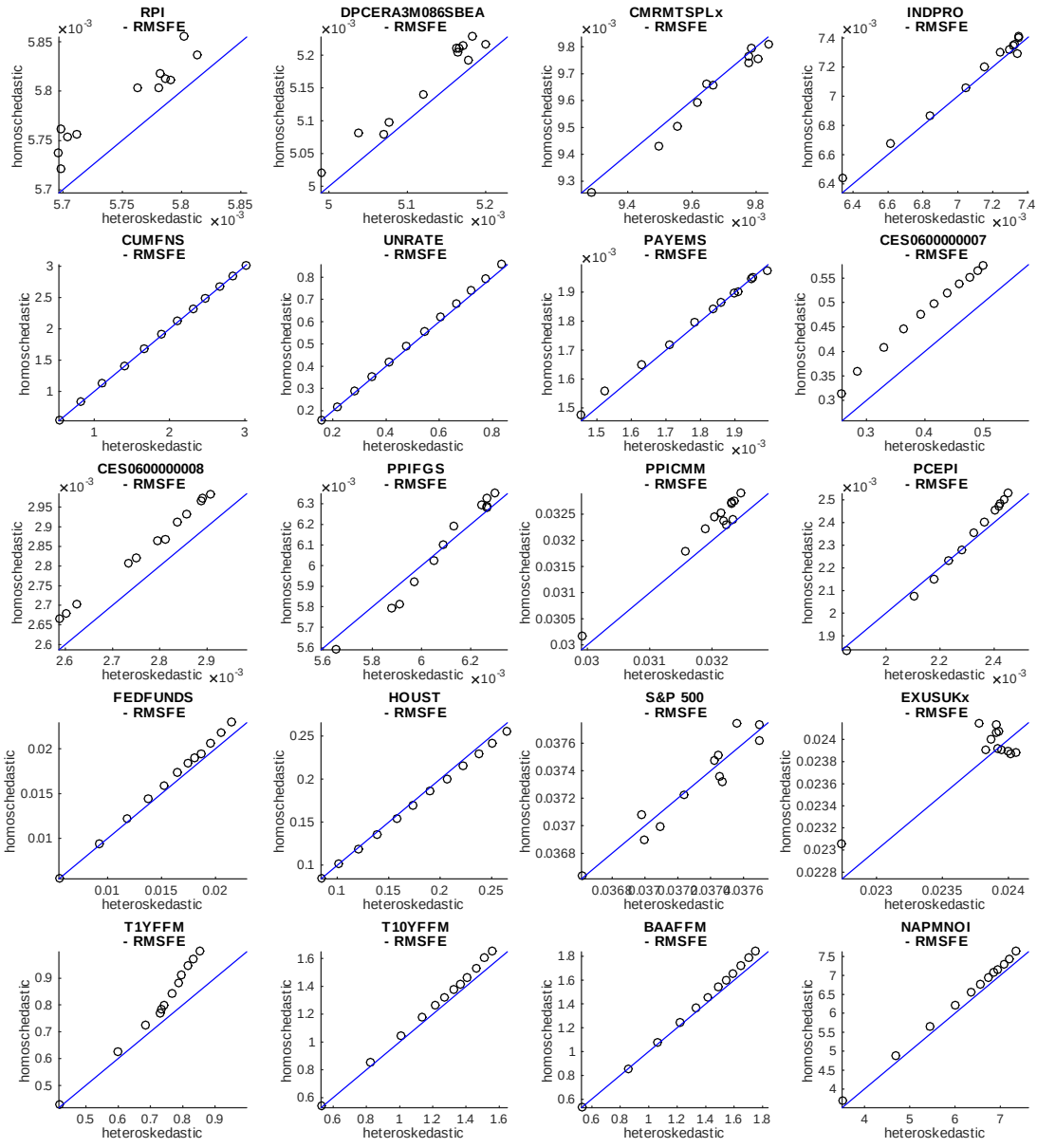
PCA of the variance matrix of the shocks to volatilities



Score comparison: homoskedastic model (y axis) vs heteroskedastic model (x axis)



RMSFE comparison: homoskedastic model (y axis) vs heteroskedastic model (x axis)



Conclusions

- The assumptions of conjugacy and homoskedasticity in a VARs are hardly defendable, but a more general specification is only manageable with a small cross-section.
- We have proposed a new estimation method VARs with **non-conjugate priors** and **drifting volatilities** which can be applied with **large** models
- 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 a single needs to be slightly modified with a few lines of code.
- 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.

Prior dependence

- We assumed that the prior variance was diagonal. This can be relaxed.
- With a prior dependent across equations, the general form of the posterior can be obtained using the triangularization also on the joint prior distribution, and is:

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

with

$$\begin{aligned}
\bar{\mu}_{\Pi^{\{j|1:j-1\}}} &= \overline{\Omega}_{\Pi^{\{j|1:j-1\}}} \left\{ \sum_{t=1}^{T} X_{j,t} h_{j,t}^{-1} y_{j,t}^{*\prime} + \underline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} \underline{\mu}_{\Pi^{\{j|1:j-1\}}} \right\} \\
\overline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} &= \underline{\Omega}^{-1}_{\Pi^{\{j|1:j-1\}}} + \sum_{t=1}^{T} X_{j,t} h_{j,t}^{-1} X_{j,t}^{\prime},
\end{aligned}$$

where $\underline{\mu}_{\Pi^{\{j|1:j-1\}}}$ and $\underline{\Omega}_{\Pi^{\{j|1:j-1\}}}$ are 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 implied by the joint prior specification.

- The moments of $\Pi^{\{j\}}|\Pi^{\{1:j-1\}}$ can be found recursively from the joint prior

Model size, stochastic volatility, and forecasting

- Pseudo out of sample exercise performed recursively, starting with the estimation sample 1960:3 to 1970:2 and ending with 1960:3 to 2014:5.
- We consider four models.
  1. A small homoskedastic VAR including the growth rate of industrial production ($\Delta \ln IP$), the inflation rate based on consumption expenditures ($\Delta \ln PECEPI$) and the effective Federal Funds Rate (FFR).
  2. A large (20 variables) homoskedastic VAR along the lines of Carriero, Clark, and Marcellino (2015), Giannone, Lenza, and Primiceri (2015), and Koop (2013).
  3. A small VAR with time variation in volatilities along the lines of Clark (2011), Cogley and Sargent (2005) and Primiceri (2005).
  4. The fourth model includes both time variation in the volatilities and a large (20 variables) information set.

Forecasting

- Direct effects:
  - The use of a larger dataset improves point forecasts via a better specification of the conditional means.
  - The inclusion of time variation in volatilities improves density forecasts via a better modelling of error variances,
- Interactions:
  - A better point forecast improves the density forecast as well, by centering the predictive density around a more reliable mean
  - Time varying volatilities improve the point forecasts 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