European Economic Review 148 (2022) 104241

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

A new algorithm for structural restrictions in Bayesian vector autoregressions☆

Dimitris Korobilis *

University of Glasgow, UK

ARTICLE INFO

JEL classification:
C11
C13
C15
C22
C52
C53
C61

Keywords:
Gibbs sampling
Factor model decomposition
Large VAR
Sign restrictions

ABSTRACT

A comprehensive methodology for inference in vector autoregressions (VARs) using sign and other structural restrictions is developed. The reduced-form VAR disturbances are driven by a few common factors and structural identification restrictions can be incorporated in their loadings in the form of parametric restrictions. A Gibbs sampler is derived that allows for reduced-form parameters and structural restrictions to be sampled efficiently in one step. A key benefit of the proposed approach is that it allows for treating parameter estimation and structural inference as a joint problem. An additional benefit is that the methodology can scale to large VARs with multiple shocks, and it can be extended to accommodate non-linearities, asymmetries, and numerous other interesting empirical features. The excellent properties of the new algorithm for inference are explored using synthetic data experiments, and by revisiting the role of financial factors in economic fluctuations using identification based on sign restrictions.

1. Introduction

This paper proposes a new Bayesian Markov chain Monte Carlo (MCMC) algorithm for joint estimation of parameters of reduced-form vector autoregressions (VARs) and associated sign restrictions for structural identification. The main idea is to allow the reduced-form VAR disturbances to have a static factor model structure. By doing so, sign and other restrictions can be incorporated via straightforward parametric prior distributions, and the factors can be interpreted as the structural VAR (SVAR) disturbances. A new, computationally efficient, algorithm is able to jointly sample VAR parameters and identification restrictions. The implication of this feature is that the parameter estimates and the fit of the VAR depend on, and interact with, the identification restrictions the researcher has in mind. Existing reduced-form VAR approaches typically follow a two-step procedure in which an estimate of the VAR covariance matrix is obtained in the first step, and some identification scheme that seems plausible to the researcher is imposed in a second step.[1] In the proposed modeling approach, different identification schemes result to different VAR parameter estimates

☆ I would like to thank without implicating Christiane Baumeister, Martin Bruns, Fabio Canova, Filippo Ferroni, Luca Gambetti, Toru Kitagawa, Gary Koop, Michele Lenza, Laura Liu, Emanuel Mönch, Alberto Musso, Serena Ng, Michele Piffer, Davide Pettenuzzo, Francesco Ravazzolo, Frank Schorfheide, Maximilian Schröder, Christian Schumacher, Leif Anders Thorsrud, John Tsoukalas and Harald Uhlig, as well as seminar and conference participants, for useful discussions and comments. Any remaining errors should solely be attributed to the author.

MATLAB code that replicates the Monte Carlo and empirical results of this paper is available on https://sites.google.com/site/dimitriskorobilis/.

* Correspondence to: Adam Smith Business School, University of Glasgow, 40 University Avenue, Glasgow, G12 8QQ, UK.
E-mail address: Dimitris.Korobilis@glasgow.ac.uk.

[1] Consider a VAR covariance matrix estimate $\hat{\boldsymbol{\Omega}}$, structural identification simply boils down to finding a matrix $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}' = \hat{\boldsymbol{\Omega}}$. There are infinite such matrices that satisfy this relationship, therefore, it is required to impose some zero or other restrictions on $\mathbf{A}$. However, these restrictions can never be (statistically) tested since $\hat{\boldsymbol{\Omega}}$ is fixed and the data likelihood remains unchanged no matter what the researcher thinks restrictions in $\mathbf{A}$ should be.

https://doi.org/10.1016/j.euroecorev.2022.104241
Received 31 March 2021; Received in revised form 13 June 2022; Accepted 8 July 2022
Available online 17 July 2022

and general model fit. The benefit of the new approach is that the researcher can treat parameter estimation and identification as a joint problem, which is a very advantageous approach towards inference, due to the fact that either the "true" VAR parameters or the "true" structural restrictions are never known. By extracting unobserved factors from VAR disturbances, sign and zero structural restrictions become, respectively, inequality and zero parametric restrictions in the associated factor loadings matrix.

The sign restrictions approach to identification has become very popular in applied work compared to traditional identification methods such as exclusion restrictions; see Kilian and Lütkepohl (2017) for a detailed review of this literature. The main feature of popular Bayesian algorithms for inference in sign restrictions, such as Rubio-Ramírez et al. (2010), is that they rely on rejection sampling schemes (also known as *accept/reject algorithms*) in order to search for matrices that satisfy the desired restrictions. If restrictions are tight, as it would be the case in models with many variables and many shocks, rejection sampling results in constantly rejecting draws. In contrast, the Gibbs sampler proposed in this paper allows to sample contemporaneous structural matrices from their conditional posterior and these samples are always accepted. The benefits of the new approach are demonstrated by revisiting the empirical results in Furlanetto et al. (2019) by using a single 15-variable VAR with many shocks, instead of many 5-variable VARs for identifying a few shocks at a time (which is what these authors do due to the computational constraints of the Rubio-Ramírez et al. (2010) algorithm they adopt). As a rough indication of the computational efficiency of the new algorithm, I find that obtaining 5000 uncorrelated samples from the benchmark six-variable VAR of Furlanetto et al. (2019) using the new algorithm takes less than five minutes; using the original (Rubio-Ramírez et al., 2010) algorithm that Furlanetto et al. (2019) adopted in order to produce their results, it takes roughly four hours to obtain 2000 draws that satisfy the same restrictions. Empirically, using the same set of sign restrictions, the two algorithms produce comparable results as measured by shapes and magnitudes of impulse response functions.

The new algorithm for structural inference in VARs shares some similarities, from a computational perspective, with the SVAR approach of Baumeister and Hamilton (2015). These authors estimate a joint model for structural restrictions and parameter estimation. However, the need to derive a reasonably simple algorithm for inference, means that these authors integrate out autoregressive and variance parameters from the joint posterior of parameters and structural restrictions using natural conjugate priors. The result is a Metropolis–Hastings algorithm that is also of the accept–reject form and cannot scale up easily to very high dimensions.[2] Most importantly, the adoption of a natural conjugate prior means that it is not possible to extend the (Baumeister and Hamilton, 2015) methods with empirically relevant time series features, such as heteroskedasticity or structural breaks. In contrast, the algorithm proposed here builds on a standard reduced-form Bayesian VAR that is easy to work with and extend to large dimensions, nonlinear or asymmetric shocks, nonlinear parameters (e.g. stochastic volatility), and numerous other interesting features; see Koop and Korobilis (2010) for a thorough review of Bayesian tools for inference in reduced-form VARs.

The idea to decompose the VAR disturbances into common factors is also related to numerous other modeling approaches. Gorodnichenko (2005) specified an identical VAR model with reduced-rank decomposition of the disturbance term. The purpose of specifying the VAR that way was to replace standard block diagonal restrictions in VARs (Bernanke and Blinder, 1992) with a more parsimonious identification scheme that imposes less (possibly unreasonable) zero restrictions. More recently, Matthes and Schwartzman (2019) specify a closely related VAR model in order to identify the structural impact of sectoral dynamics on GDP. Their identification is via a factor structure on the residuals that has the additional assumption of allowing for correlation within industries but no correlation across industries.

In a different strand of the VAR literature, Stock and Watson (2005) specify a more general factor-augmented VAR (FAVAR) and discuss in detail how various identification schemes fit in this setting. They note (Stock and Watson, 2005, Section 3.5) that the sign restrictions identification scheme proposed by Uhlig (2005) also fits the FAVAR framework. An application of this idea can be found in Ahmadi and Uhlig (2015). From a modeling point of view, the model I propose in this paper can be viewed as a special case of the Ahmadi and Uhlig (2015) FAVAR model. However, the specification I use has completely different implications both algorithmically and in terms of inference. Ahmadi and Uhlig (2015) project a large vector of observable macroeconomic variables into a smaller vector of factors and they model VAR dynamics only on these factors. This means that there is some loss of information (not all macro variables are explained well by the factors) and the statistical fit of the factors determines the contribution of each structural shock on each macroeconomic variable. Additionally, the autoregressive dynamics of the large macro dataset is represented only by the autoregressive dynamics of the smaller vector of factors. This modeling choice means that, inevitably, the FAVAR is unable to capture richer patterns of propagation of structural shocks to observed macroeconomic variables. In contrast, in this paper all observable macroeconomic variables are endogenous in the VAR and the sole role of the factors is to represent structural shocks. Additionally, the algorithm derived here is computationally simpler as it relies on posterior formulas for linear regression models, instead of building on more demanding simulation smoothing techniques, as is the case with the FAVAR (see Ahmadi and Uhlig (2015) and Bernanke et al. (2005)).

The next Section introduces the new methodology and associated Gibbs sampler algorithm for inference, and it outlines the key components that help speed up and stabilize (numerically) posterior sampling in high dimensions. In Section 3 I undertake several important exercises using synthetic datasets, in order to test both the computational features of the new algorithm as well as shed light on how joint inference on parameters and structural restriction is implemented. In Section 4 the algorithm is applied to the issue of measuring the impact of a financial shock to the macroeconomy. Section 5 concludes the paper.

[2] To be exact, if a candidate sample is not accepted as a sample from the true posterior, then the immediately previous accepted sample values are used. Bruns and Piffer (2019) propose a more efficient Dynamic Striated Metropolis Hastings algorithm that builds on importance sampling proposals. While this algorithm is appropriate for high-dimensional models, in the case of the VAR it will still hit a computational bottleneck at much lower dimensions than the Gibbs sampler proposed in this paper.

2. A new methodology for sign restrictions in VARs

The starting point is the reduced-form vector autoregression

$$\mathbf{y}_t = \mathbf{\Phi}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{y}_t$ is a $(n \times 1)$ vector of observed variables, $\mathbf{x}_t = \left(1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p}\right)'$ a $(k \times 1)$ vector (with $k = np+1$) containing a constant and p lags of $\mathbf{y}$, $\mathbf{\Phi}$ is an $(n \times k)$ matrix of coefficients, and $\boldsymbol{\varepsilon}_t$ a $(n \times 1)$ vector of disturbances distributed as $N\left(\mathbf{0}_{n\times 1}, \mathbf{\Omega}\right)$ with $\mathbf{\Omega}$ an $n \times n$ covariance matrix. The structural VAR (SVAR) form associated with the reduced-form model in (1) is

$$\mathbf{A}\mathbf{y}_t = \mathbf{B}\mathbf{x}_t + \boldsymbol{u}_t, \tag{2}$$

where $\mathbf{B} = \mathbf{A}\mathbf{\Phi}$, $\boldsymbol{u}_t = \mathbf{A}\boldsymbol{\varepsilon}_t$ and $cov(\boldsymbol{u}_t) = \mathbf{D}$, with $\mathbf{D}$ an $n \times n$ diagonal matrix which, sometimes, is normalized to be the identity matrix. The SVAR form can be obtained by means of a decomposition of the reduced-form covariance matrix of the form $\mathbf{A}\mathbf{\Omega}\mathbf{A}' = \mathbf{D}$ where both sides of Eq. (1) are left-multiplied with the $n \times n$ matrix $\mathbf{A}$. This decomposition is unique when $\mathbf{A}$ is a lower triangular matrix (typically with a unit diagonal, unless we do the normalization $\mathbf{D} = \mathbf{I}$), but it has infinite solutions for a full matrix $\mathbf{A}$.

2.1. VARs driven by a few, common shocks

I begin by building on fundamental ideas introduced in the factor model literature, as applied to empirical problems in macroeconomics: a few common forces (which in a structural setting we desire to identify as "primitive shocks"; see Ramey, 2016) are driving the set of reduced-form shocks in the system of n endogenous variables. In order to materialize this idea, the reduced-form VAR disturbances in Eq. (1) are decomposed using the following static factor model specification

$$\boldsymbol{\varepsilon}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{v}_t, \tag{3}$$

where $\mathbf{\Lambda}$ is an $n \times r$ matrix of factor loadings, $\mathbf{f}_t$ is an $r \times 1$ vector of factors, and $\mathbf{v}_t$ is an $n \times 1$ vector of idiosyncratic shocks. While the n shocks in $\boldsymbol{\varepsilon}_t$ are decomposed into $r + n$ shocks, only the r common shocks in $\mathbf{f}_t$ are considered structural and the n shocks in $\mathbf{v}_t$ are simply nuisance shocks (e.g. due to measurement or expectations error, incomplete information etc.). The assumption here is that n is large and that $r \leq n$, and not necessarily $r \ll n$, as is typically assumed in the factor literature. In line with the *exact factor model* literature, let $\mathbf{v}_t \overset{\text{i.i.d}}{\sim} N\left(\mathbf{0}_{n\times 1}, \mathbf{\Sigma}\right)$, with $\mathbf{\Sigma}$ an $n \times n$ diagonal matrix. Additionally, let $\mathbf{f}_t \sim N\left(\mathbf{0}_{r\times 1}, \mathbf{I}_r\right)$, such that the conditional covariance matrix of $\boldsymbol{\varepsilon}_t$ is now of the form

$$cov\left(\boldsymbol{\varepsilon}_t | \mathbf{\Lambda}, \mathbf{\Sigma}\right) = \mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Sigma}. \tag{4}$$

This factor model decomposition of $\mathbf{\Omega}$ reveals that, as long as $\mathbf{\Sigma}$ is diagonal, identification via sign restrictions can be achieved by imposing the desired signs on $\mathbf{\Lambda}$. Similarly, zero restrictions simply correspond to setting the respective elements of $\mathbf{\Lambda}$ to zero.[3] To see this, consider a reduced-rank SVAR representation of this model, which can be obtained by left-multiplying the reduced-form VAR model given by Eqs. (1)–(3) with the generalized inverse of $\mathbf{\Lambda}$, as follows:

$$\mathbf{y}_t = \mathbf{\Phi}\mathbf{x}_t + \mathbf{\Lambda}\mathbf{f}_t + \mathbf{v}_t \tag{5}$$

$$\left(\mathbf{\Lambda}'\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{y}_t = \left(\mathbf{\Lambda}'\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{\Phi}\mathbf{x}_t + \mathbf{f}_t + \left(\mathbf{\Lambda}'\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{v}_t \tag{6}$$

$$\mathbf{A}_1\mathbf{y}_t = \mathbf{B}_1\mathbf{x}_t + \mathbf{f}_t + \left(\mathbf{\Lambda}'\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{v}_t, \Rightarrow \tag{7}$$

$$\mathbf{f}_t \approx \mathbf{A}_1\mathbf{y}_t - \mathbf{B}_1\mathbf{x}_t. \tag{8}$$

In the equation above, the SVAR matrix $\mathbf{A}_1$ is equivalent to the generalized inverse $\left(\mathbf{\Lambda}'\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'$. While $\mathbf{\Lambda}$ is not observed, assume that a consistent estimator of this parameter exists. Given that in the exact factor model formulation the $\mathbf{v}_t$ are uncorrelated, the Central Limit Theorem in Bai (2003) suggests that for each t and for $n \to \infty$ we have $\left(\mathbf{\Lambda}'\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{v}_t \to 0$ making this term asymptotically negligible. Therefore, it is justified to view $\mathbf{v}_t$ as a residual or a noise shock that carries no structural interpretation. At the same time, f_t can be interpreted as a projection of the SVAR structural shocks $\mathbf{u}_t$ into $\mathbb{R}^r$.

Impulse response functions (IRFs) can be obtained via the vector moving average (VMA) representation of the VAR. In the particular case with $p = 1$ lags (for notational simplicity) the VMA form becomes

$$\mathbf{y}_t = \phi_0 + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \Rightarrow \tag{9}$$

$$\mathbf{y}_t = \mu + \sum_{i=0}^{\infty}\mathbf{\Psi}_i\boldsymbol{\varepsilon}_t, \Rightarrow \tag{10}$$

$$\mathbf{y}_t = \mu + \sum_{i=0}^{\infty}\mathbf{\Psi}_i\mathbf{\Lambda}\mathbf{f}_t + \sum_{i=0}^{\infty}\mathbf{\Psi}_i\mathbf{v}_t, \tag{11}$$

[3] If desired, several other restrictions can be incorporated in a straightforward way, such as ranking restrictions (Amir-Ahmadi and Drautzburg, 2021), restrictions on elasticities, or other restrictions on magnitudes of shocks. For example, if for shock j variable i should react with a larger magnitude than variable k, then we get the restriction $\mathbf{\Lambda}_{ij} > \mathbf{\Lambda}_{kj}$ which is fairly simple to incorporate within an MCMC sampling setting. The key early reference for inference in the Bayesian regression model with general inequality constraints is Geweke (1996).

where $\mu = \left(I - \mathbf{\Phi}_1 L\right)^{-1} \phi_0$ with L is the lag operator, and $\mathbf{\Psi}_i = \sum_{j=1}^{i} \mathbf{\Psi}_{j-i} \mathbf{\Phi}_1$ with $\mathbf{\Psi}_0 = \mathbf{I}$ (see Kilian and Lütkepohl, 2017, Section 2.2.2). The impulse response on impact is

$$\frac{\partial \mathbf{y}_t}{\partial \mathbf{u}_t} \approx \frac{\partial \mathbf{y}_t}{\partial \mathbf{f}_t} = \mathbf{\Psi}_0 \mathbf{\Lambda} = \mathbf{\Lambda}, \tag{12}$$

showing that parametric restrictions in $\mathbf{\Lambda}$ correspond to structural restrictions on impact IRFs. This result is true even when considering the effect of $\mathbf{v}_t$, since this term has a diagonal covariance matrix and does not affect contemporaneous relationships in the variables $\mathbf{y}_t$.

Finally, similar to the algorithm in Baumeister and Hamilton (2015), the proposed specification is efficient only for static sign restrictions. This is because dynamic restrictions are nonlinear and cannot be represented (in a straightforward way) as equivalent parametric inequality restrictions in the linear VAR parameters. Therefore, dynamic restrictions would require to turn to less efficient sampling schemes similar to Arias et al. (2018). In practice, there is little consensus in economic theory about the signs of structural impulse responses at longer horizons (see for example Canova and Paustian, 2011), and for that reason the vast majority of empirical papers impose restrictions only on impact (Kilian and Lütkepohl, 2017). Nevertheless, as Uhlig (2017) notes, it can be quite useful to have the option to impose sign restrictions in the longer-run responses of macroeconomic variables to shocks. Within the context of the proposed methodology, this issue can be addressed if Eq. (3) is specified as a dynamic factor model instead of a static factor model. However, this more general modeling assumption would require to rely on filtering and smoothing sampling steps that would, in turn, lead to a completely different estimation algorithm compared to the algorithm presented in this paper. As a result, I exclusively focus here on impact (contemporaneous) structural restrictions in small and large VARs. Extending to the case of long-horizon sign restrictions is conceptually feasible, but it is left for future research.

2.2. Identification

The previous discussion established that the factor decomposition of the VAR disturbances projects the n shocks into r structural plus n nuisance shocks. Therefore, the first identification issue relates to being able to separate the common component ($\mathbf{\Lambda}\mathbf{f}_t$) from the idiosyncratic one. Notice that the original VAR covariance matrix $\mathbf{\Omega}$ has $n(n+1)/2$ free elements, while the right-hand side of the factor decomposition in (4) has $nr+n$ free parameters. Therefore, the first condition is that $n(n+1)/2 \geq nr+n$ or that $r \leq (n-1)/2$, which implies that the common component will always be identified even if the factors (structural shocks) are not identified. This condition implies that in a 19-variable VAR a reasonable number of nine factors/shocks can be estimated. Next, restrictions are required for uniquely identifying the factors, which are also structural shocks. Following Anderson and Rubin (1956) an additional $r(r-1)/2$ restrictions are needed in order to deal with the *rotation problem*. This is due to the fact that rotating $\mathbf{\Lambda}$ and $\mathbf{f}_t$ using an orthogonal matrix $\mathbf{P}$ leads to an observationally equivalent solution, that is,

$$\mathbf{\Lambda}\mathbf{f}_t = \mathbf{\Lambda}\mathbf{P}\mathbf{P}'\mathbf{f}_t = \widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{f}}_t, \tag{13}$$

where $\widetilde{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{P}$ and $\widetilde{\mathbf{f}}_t = \mathbf{P}'\mathbf{f}_t$.

From an estimation perspective, the above equation shows that sampling of unique values of the factors in Eq. (3) cannot be achieved without these additional $r(r-1)/2$ restrictions. From a structural perspective, this condition shows that the same number of restrictions is required for identification of the structural factors (shocks) in the VMA form of Eq. (11). Consequently, under the assumption that $\mathbf{f}_t \sim N(\mathbf{0}, \mathbf{I})$, placing restrictions on $\mathbf{\Lambda}$ ensures both unique estimation of factors and identification of the structural model. For instance, Anderson and Rubin (1956) show that setting to zero all elements of $\mathbf{\Lambda}$ above the main diagonal achieves the $r(r-1)/2$ restrictions required for identification of the factors. Additionally, the diagonal elements of $\mathbf{\Lambda}$ can be normalized to be nonnegative, such that the sign of the factors is also always identified. However, as shown in detail in Stock and Watson (2005) numerous other identifying assumptions can be used in structural factor models.

In this paper interest lies in structural identification via sign (and possibly some zero) restrictions. However, many other restrictions can be incorporated in a straightforward way. For example, a researcher may want to impose that the impact response of variable i to shock j is lower in magnitude than the response of variable k to the same shock. Due to the fact that impact restrictions on the IRFs are equivalent to imposing restrictions to $\mathbf{\Lambda}$ (see Eq. (12)), this magnitude restriction can be represented as $\mathbf{\Lambda}_{ij} < \mathbf{\Lambda}_{kj}$. In the proposed model, a large class of desired structural restrictions can be represented using parametric inequalities that are imposed upon estimation of $\mathbf{\Lambda}$. Following Geweke (1996), these parametric inequality restrictions can be sampled efficiently using the Gibbs sampler. The next subsection derives such a Gibbs sampler algorithm, with particular focus on ensuring computational efficiency in high-dimensions.

2.3. Gibbs sampler for sign and zero restrictions in reduced-form VARs

Posterior sampling in the reduced-form VAR with factor structure in the residuals is straightforward due to the fact that posterior conditional distributions have standard forms. To see this, write the model using a single equation for convenience

$$\mathbf{y}_t = \mathbf{\Phi}\mathbf{x}_t + \mathbf{\Lambda}\mathbf{f}_t + \mathbf{v}_t. \tag{14}$$

Assume that all sign and zero restrictions in $\mathbf{\Lambda}$ are collected into a matrix $\mathbf{S}$, with entries $+1$ for positive signs, -1 for negative signs, 0 for zero restrictions, and a missing value for no restriction (this case is denoted in this paper using the symbol *NA*, and in the code using the MATLAB value NaN). The priors for the VAR parameters are of the form

$$\boldsymbol{\phi}_i \equiv vec\left(\mathbf{\Phi}_i\right) \sim N_k\left(\mathbf{0}, \underline{\mathbf{V}}_i\right), \tag{15}$$

$$\mathbf{f}_t \sim N_r(\mathbf{0}, \mathbf{I}), \tag{16}$$

$$\mathbf{\Lambda}_{ij} \sim \begin{cases} N\left(0, \underline{h}_{ij}\right) I(\Lambda_{ij} > 0), & \text{if } S_{ij} = 1, \\ N\left(0, \underline{h}_{ij}\right) I(\Lambda_{ij} < 0), & \text{if } S_{ij} = -1, \\ \delta_0\left(\mathbf{\Lambda}_{ij}\right), & \text{if } S_{ij} = 0, \\ N\left(0, \underline{h}_{ij}\right), & \text{otherwise,} \end{cases} \tag{17}$$

$$\sigma_i^2 \sim inv - Gamma\left(\underline{\rho}_i, \underline{\kappa}_i\right), \tag{18}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, r$, where $\mathbf{\Phi}_i$ is the ith row of $\mathbf{\Phi}$, σ_i^2 is the ith diagonal element of the matrix $\mathbf{\Sigma}$, and $\delta_0\left(\mathbf{\Lambda}_{ij}\right)$ is the Dirac delta function for $\mathbf{\Lambda}_{ij}$ at zero (i.e. a point mass function with all mass concentrated at zero).

The joint posterior of the parameters is the distribution $p(\boldsymbol{\Phi}, \boldsymbol{\Lambda}, \boldsymbol{F}, \boldsymbol{\Sigma}|\boldsymbol{y}) \equiv p\left(\{\boldsymbol{\phi}_i\}_{i=1}^n, \{\mathbf{\Lambda}_i\}_{i=1}^n, \{\mathbf{f}_t\}_{t=1}^T, \{\sigma_i\}_{i=1}^n|\mathbf{y}\right)$, which by Bayes theorem is the product of the normal likelihood function implied by Eq. (14) and the prior distributions presented above. This product is a complicated function making sampling from the joint posterior infeasible. However, the conditional posteriors are tractable and trivial to derive in this linear model. Therefore, Bayesian inference in this VAR breaks down to sequentially sampling from the following conditional posterior distributions

Factor-based structural restrictions (FSR) algorithm

1. Sample $\boldsymbol{\phi}_i$ for $i = 1, \ldots, n$ from

$$\boldsymbol{\phi}_i|\mathbf{\Sigma}, \mathbf{\Lambda}, \mathbf{f}, \mathbf{y} \sim N_k\left(\overline{\mathbf{V}}_i\left(\sum_{t=1}^T \sigma_i^{-2}\mathbf{x}_t'\widetilde{\mathbf{y}}_{it}\right), \overline{\mathbf{V}}_i\right), \tag{19}$$

where $\widetilde{\mathbf{y}}_{it} = \mathbf{y}_{it} - \mathbf{\Lambda}_i\mathbf{f}_t$ and $\overline{\mathbf{V}}_i^{-1} = \left(\underline{\mathbf{V}}_i^{-1} + \sum_{t=1}^T \sigma_i^{-2}\mathbf{x}_t'\mathbf{x}_t\right)$.

2. Sample $\mathbf{\Lambda}_i$ for $i = 1, \ldots, n$ from

$$\mathbf{\Lambda}_i|\mathbf{\Phi}, \mathbf{\Sigma}, \mathbf{f}, \mathbf{y} \sim MTN_{\mathbf{a}<vec(\mathbf{\Lambda})<\mathbf{b}}\left(\overline{\mathbf{H}}_i\left(\sum_{t=1}^T \sigma_i^{-2}\mathbf{f}_t'\widehat{\mathbf{y}}_{it}\right), \overline{\mathbf{H}}_i\right), \tag{20}$$

where $\widehat{\mathbf{y}}_{it} \equiv \boldsymbol{\varepsilon}_{it} = \mathbf{y}_{it} - \phi_i\mathbf{x}_t$, $\overline{\mathbf{H}}_i^{-1} = \left(\underline{\mathbf{H}}_i^{-1} + \sum_{t=1}^T \sigma_i^{-2}\mathbf{f}_t'\mathbf{f}_t\right)$, and $\underline{\mathbf{H}}_i = diag\left(\underline{h}_{i1}, \ldots, \underline{h}_{ir}\right)$. Here we define $MTN(\bullet)$ to be the multivariate truncated Normal distribution, and $\mathbf{a}, \mathbf{b}$ are the vectors indicating the truncation points, with ijth element:

$$(\mathbf{a}_{ij}, \mathbf{b}_{ij}) = \begin{cases} (-\infty, 0) & \text{if } S_{ij} = -1, \\ (0, \infty) & \text{if } S_{ij} = 1, \\ (0, 0) & \text{if } S_{ij} = 0, \\ (-\infty, \infty) & \text{otherwise,} \end{cases} \tag{21}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, r$.

3. Sample $\mathbf{f}_t$ for $t = 1, \ldots, T$ from

$$\mathbf{f}_t|\mathbf{\Lambda}, \mathbf{\Sigma}, \mathbf{\Phi}, \mathbf{y} \sim N\left(\overline{\mathbf{G}}\left(\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\widehat{\mathbf{y}}_t\right), \overline{\mathbf{G}}\right), \tag{22}$$

where $\overline{\mathbf{G}}^{-1} = \left(\mathbf{I}_r + \mathbf{\Lambda}'\mathbf{\Sigma}\mathbf{\Lambda}\right)$.

4. Sample σ_i^2 for $i = 1, \ldots, n$ from

$$\sigma_i^2|\mathbf{\Lambda}, \mathbf{f}, \mathbf{\Phi}, \mathbf{y} \sim inv - Gamma\left(\frac{T}{2} + \underline{\rho}_i, \left[\underline{\kappa}_i^{-1} + \sum_{t=1}^T \left(\mathbf{y}_{it} - \phi_i\mathbf{x}_t - \mathbf{\Lambda}_i\mathbf{f}_t\right)'\left(\mathbf{y}_{it} - \phi_i\mathbf{x}_t - \mathbf{\Lambda}_i\mathbf{f}_t\right)\right]^{-1}\right) \tag{23}$$

Step 1 is efficient because autoregressive coefficients can be sampled equation-by-equation. Further speed enhancements can be achieved by using the sampling methodology of Bhattacharya et al. (2016) in the case where the prior covariance matrices $\underline{\mathbf{V}}_i$ are diagonal. This is the case here, as $\underline{\mathbf{V}}_i$ is diagonal and its elements follow the horseshoe hierarchical structure of Carvalho et al. (2010) which has the form

$$\boldsymbol{\phi}_i|\sigma_i^2\tau_i^2\mathbf{\Psi}_i \sim N_k\left(\mathbf{0}, \underline{\mathbf{V}}_i\right), \tag{24}$$

$$\underline{\mathbf{V}}_{i,(jj)} = \sigma_i^2\tau_i^2\psi_{i,j}^2, \tag{25}$$

$$\psi_{i,j} \sim Cauchy^+(0, 1), \tag{26}$$

$$\tau_i \sim Cauchy^+(0, 1). \tag{27}$$

This prior belongs to the class of *local–global* shrinkage priors, that is a prior that penalizes the likelihood and shrinks coefficients towards zero. In this prior, $\psi_{i,j}$ is the local shrinkage parameter for each scalar coefficient $\boldsymbol{\phi}_{i,j}$ while τ_i is the global shrinkage

parameter pertaining to all coefficients in equation i. Unlike the popular (in macroeconomic VARs) Minnesota prior that typically requires subjective tuning (Koop and Korobilis, 2010), the horseshoe prior is tuning-free as the local and global shrinkage parameters have their own distributions and are, thus, updated by information in the data. The horseshoe prior has established posterior consistency properties when used in a variety of high-dimensional settings[4] making it an appropriate choice for penalized estimation in both smaller and higher dimensional VARs. Sampling from the truncated Normal posteriors in step 2 of the algorithm above can be done using the recent contribution of Botev (2017).[5] Finally, sampling of the factors $\mathbf{f}_t$ for each $t = 1, \dots, T$ is fairly fast for monthly or quarterly macroeconomic data, as is sampling of the scalar parameters σ_i^2 for each $i = 1, \dots, n$. Computational details and further discussion on the excellent properties of the horseshoe prior are provided in the online Appendix.

Given that the horseshoe prior requires no subjective tuning, there are only a handful of prior hyperparameters in the whole VAR that need to be elicited by the researcher. The parameter affecting primarily structural identification, is the choice of the prior variance $\underline{h}_{ij}$, regardless of whether the associated loading parameter $\mathbf{\Lambda}_{ij}$ should be truncated (sign restricted) or not. Due to the fact that the coefficients $\mathbf{\Lambda}$ enter the VAR in a linear way, it turns out that a fairly large value of $\underline{h}_{ij}$ implies a diffuse prior that does not impact the posterior asymptotically. Therefore, for the remainder of this paper, I set $\underline{h}_{ij} = 100$ which is a fairly diffuse choice for the typical scale of variables encountered in macroeconomic VARs. Finally, I set $\rho_i = 1$ and $\kappa_i = 0.01$ for all i, leading to a prior mean of 0.16 and substantially large prior variance, reflecting the belief that the scale of the idiosyncratic variances σ_i^2 should be small and most of variability in the VAR disturbances should come from the common component $\mathbf{\Lambda}\mathbf{f}_t$. Model fit in this VAR can be assessed with the Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002), due to its simplicity of implementation. The formula and justification for the use of the DIC is provided also in the online Appendix. What suffices to remember is that the DIC has the same interpretation as any other information criterion, that is, lower values signify better fit. The DIC can be used to test any kind of parametric restrictions that are of interest, whether it pertains to lag length selection, number of structural shocks, linearities vs nonlinearities, and so on.

The most important computational aspect of the new algorithm is that samples from the restricted $\mathbf{\Lambda}$ matrices are always accepted, making it very efficient in high-dimensions. This feature is in contrast with a large class of accept/reject algorithms used especially for sign restrictions in VARs; see for example the Bayesian algorithms of Rubio-Ramírez et al. (2010) and Baumeister and Hamilton (2015) as well as the algorithm of Ouliaris and Pagan (2016). In Rubio-Ramírez et al. (2010), for example, one has to first obtain posterior samples from the VAR covariance matrix $\mathbf{\Omega}$ and then rotate its Cholesky factor $\mathbf{P}$ using randomly generated orthogonal matrices $\mathbf{Q}$. If the random rotation $\mathbf{H} = \mathbf{PQ}$ satisfies the required sign restrictions then $\mathbf{H}$ is a draw from the desired matrix of contemporaneous structural relationships. Inevitably, such an accept/reject scheme is deemed to fail in high-dimensions, when the desired restrictions may be so tight that no sample of $\mathbf{H}$ can be accepted by random chance, see also the discussion in Section 13.6.4 of Kilian and Lütkepohl (2017). Many authors express the belief that when the accept/reject algorithm results to a high rejection rate, this is evidence that identification is sharp.[6] However, this premise is not testable in a statistical sense, and can be misleading especially in high dimensions: a high rejection rate could either be because of the researcher imposing too many restrictions, or because of imposing restrictions that simply do not comply with the evidence in the data.

3. Simulation studies

3.1. Numerical evaluation of the new algorithm

In this section, the properties of the new algorithm are explored using artificially generated data. The core exercise involves generating multivariate time series from a data generating process (DGP) that fully matches equation (14), and estimating parameters and impulse response functions based on time series generated from this DGP. I first implement this experiment assuming that a correctly specified model is estimated using the artificial data. Subsequently, various cases of misspecification errors during the estimation process are considered — that is, I estimate models that do not perfectly match the correct DGP.

The DGP is of the form

$$\mathbf{y}_t = \widehat{\mathbf{\Phi}}\mathbf{x}_t + \widehat{\mathbf{\Lambda}}\mathbf{f}_t + \mathbf{v}_t, \text{ for } t = 1, \dots, \widehat{T}, \tag{28}$$

$$\mathbf{v}_t \sim N\left(\mathbf{0}, \widehat{\mathbf{\Sigma}}\right), \quad \mathbf{f}_t \sim N\left(\mathbf{0}, \mathbf{I}\right), \tag{29}$$

$$\mathbf{y}_{(-p+1):0} = \mathbf{0}, \quad p = 12, \quad r = 3. \tag{30}$$

[4] See online Appendix for more details and citations.

[5] In an ideal world one would want to sample the vector $\mathbf{\Lambda}_i$ in one step and unconditionally from the factors $\mathbf{f}_t$, in order to reduce correlation in the MCMC chain. In practice, I follow Geweke (1996) and sample each element $\mathbf{\Lambda}_{ij}$ conditional on $\mathbf{\Lambda}_{ij}$, as this conditioning allow for the algorithm to be extended more easily (e.g. if structural breaks or time-varying loadings are required). This comes at the cost of thinning the Gibbs chain by a factor of 100, that is, every 100th sample is stored in order to ensure posterior samples are uncorrelated. In all results in this paper the Gibbs chain runs for 550,000 iterations where the first 50,000 iterations are discarded and from the final 500,000 iterations I store every 100th iteration, leaving a total of 5000 samples from the parameter posterior for inference.

[6] See for example "principle 7" in Uhlig (2017) and the corresponding discussion.

Table 1
OLS estimates $\widehat{\mathbf{\Lambda}}$ used in the DGP, and sign restrictions $\mathbf{S}$ used for estimation.

Variable	(a) True parameter values			(b) Sign restrictions		
	1st shock	2nd shock	3rd shock	1st shock	2nd shock	3rd shock
Real GDP growth	1.00	−1.39	−0.87	+	−	−
GDP deflator inflation	1.42	1.00	−0.71	+	+	−
Fed funds rate	0.49	−0.28	1.00	NA	NA	+
Commodity prices	0.16	0.16	−0.45	NA	NA	−
Total reserves	−0.61	0.22	−3.48	NA	NA	NA
Nonborrowed reserves	−0.91	0.25	−3.37	NA	NA	−
Stock prices	−0.25	−0.30	−0.82	NA	NA	−
M1	−1.03	−0.48	−1.27	−	−	−
Unemployment	−0.63	0.51	0.43	−	+	+
Industrial production	1.12	−1.34	−0.87	+	−	−
Employment	0.88	−1.00	−1.01	+	−	−
CPI inflation (total)	1.44	1.01	−0.75	+	+	−
CPI inflation (core)	1.05	0.49	−1.12	+	+	−
PCE inflation (core)	1.05	0.57	−0.80	+	+	−

Notes: Panel (A) shows true parameter values used as input in the data generating process (DGP), while panel (B) shows the sign restrictions imposed during econometric estimation using each artificial dataset from the DGP. Entries in panel (B) show the restrictions imposed: + for positive sign; − for negative sign; NA for no restriction.

The DGP parameters $\widehat{\mathbf{\Phi}}$, $\widehat{\mathbf{\Lambda}}$, $\widehat{\mathbf{\Sigma}}$ are based on estimates of a VAR on real data. First, monthly data on 14 monthly macroeconomic variables are collected[7] for the US over the period 1965M1–2007M12, providing $\widehat{T} = 516$ observations.[8] At a second step, an estimate $\widehat{\mathbf{\Phi}}$ is obtained by applying OLS to an unrestricted VAR(12) estimated with these 14 observed US variables. The third step is to obtain the first $r = 3$ principal components of these OLS residuals, and store the estimate $\widehat{\mathbf{\Lambda}}$ using OLS in a regression between the VAR residuals and their principal components. Finally, the residuals from this latter regression provide the elements of the diagonal matrix $\widehat{\mathbf{\Sigma}}$, by means of equation-by-equation application of the usual least squares formula for the variance.

While it is not possible, or even interesting, to display all estimates $\widehat{\mathbf{\Phi}}$ used as input in the DGP, it is instead interesting to look at the estimates $\widehat{\mathbf{\Lambda}}$ obtained using the procedure described above. This is because both the signs and the magnitudes of the implied IRFs in the true DGP will be affected by those estimates. Panel (A) of Table 1 shows the OLS estimates, where the diagonal is normalized to be one, by dividing each element in the mth column of $\widehat{\mathbf{\Lambda}}$ with the original value of its mth element, $m = 1, 2, 3$. While this matrix is the outcome of using real data and applying simple principal components followed by OLS estimates (which carry no structural restrictions), looking at the signs of the loadings of the first three variables (output, inflation, interest rate) allows for the classification of the three pseudo-shocks as aggregate supply, aggregate demand, and monetary policy, respectively. The estimated magnitudes, of course, are not necessarily economically meaningful. For example, in the first column a shock of 1% in GDP increases inflation by 1.49%, which is probably not a representative magnitude for a true aggregate supply shock. Nevertheless, this is an exercise where the main aim is to check the numerical precision of the new algorithm, so the estimates in panel (A) of Table 1 are perfectly valid inputs for a DGP. Finally, panel (B) of Table 1 shows the sign restrictions imposed on $\mathbf{\Lambda}$, that is the matrix $\mathbf{S}$ introduced in Eq. (17). These comply with the signs imposed in the DGP, and in 11 instances no sign restrictions are imposed (these entries are denoted as *NA*). The choice of which signs are known in $\mathbf{S}$ is random, and the next subsection looks at varying assumptions about how many sign restrictions are known to the researcher.

For estimation purposes five different scenarios are assumed: one correctly specified case and four misspecified cases. These are denoted as C1–C5, and are defined as follows:

- **C1** Correctly specified model with $n = 14$ dependent variables, $p = 12$ lags, $r = 3$ shocks.
- **C2** Misspecified model with $n = 8$, using the first eight variables in Table B1 in the online Appendix. All other settings are correct, that is, $p = 12$ lags, $r = 3$ shocks.
- **C3** Misspecified model with $p = 2$ lags. All other settings are correct, that is, $n = 14$ and $r = 3$.
- **C4** Misspecified model with $r = 2$ shocks, using only the restrictions on the first two shocks in panel (B) of Table 1. All other settings are correct, that is, $n = 14$ and $p = 12$.
- **C5** Misspecified model with $r = 4$ shocks, using an additional shock.[9] All other settings are correct, that is, $n = 14$ and $p = 12$.

500 datasets of size $T = 516$ are generated and posterior mean estimates of all parameters, IRFs and DICs from all five cases above are obtained. Results presented next are based on the distribution of the posterior means over these 500 generated datasets.

[7] The variables are: (1) real GDP, (2) GDP deflator, (3) federal funds rate, (4) commodity price index, (5) total reserves, (6) nonborrowed reserves, (7) S&P 500, (8) M1, (9) unemployment rate, (10) industrial production, (11) employment, (12) CPI, (13) core CPI, (14) core PCE. More details on these variables is provided in the online Appendix.

[8] In practice, I generate $\widehat{T} + 1000$ observations and discard the first 1000 observations in order to remove dependence to the initial values of the generated time series process.

[9] This fourth shock is identified using the randomly selected vector of restrictions $s = [+, +, +, -, +, NA, NA, -, +, +, +, +, +, +]$.

Table 2
DIC values attained by the correctly specified and misspecified models estimated on artificial data.

	C1	C2	C3	C4	C5
DIC_{14} value	7393.14	n/a	27 241.13	11 859.45	9037.44
DIC_{8} value	15 300.63	17 094.28	28 630.21	27 981.48	19 269.55

Notes: DIC_{14} is the deviance information criterion applied jointly to all 14 VAR equations. DIC_{8} is the same criterion applied jointly only to the first 8 VAR equations. Case C2 does not have a DIC_{14} value because it assumes that the VAR has $n = 8$ variables.

Before evaluating precision of estimates over the Monte Carlo iterations, it is important to first evaluate general model fit using the DIC. Table 2 shows the value of the deviance information criterion attained by the estimates of the VAR model in each of the five specification cases. Because case C2 refers to a VAR with $n = 8$, it is impossible to directly compare it with the other four cases that assume $n = 14$.[10] For that reason I present two DIC metrics, a full one based on all $n = 14$ variables (with no value available for C2) and a reduced DIC which is the same formula evaluated only in the first eight VAR equations (which are common to all five cases). These are labeled in Table 2 as DIC_{14} and DIC_{8}, respectively. According to both subsets of criteria, the correctly specified estimated model, case C1, is the best one as it attains the lowest DIC value. Interestingly, the case where an additional fourth shock is incorrectly estimated (C5), does not seem to harm estimation accuracy; at least not as much as the case of estimating one less shock (C4). This result makes sense because shocks in the proposed VAR are equivalent to factors. By far the worst type of misspecification seems to be the one related to the lag-length. This is a characteristic of the VAR model rather than a "problem" with the specific algorithm or prior. As long as the true DGP has $p = 12$ important lags, estimating the VAR with $p = 2$ provides a huge loss of structurally meaningful information. In contrast, reducing the VAR from $n = 14$ (which is the truth in the DGP) to $n = 8$ as in case C2 is less harmful for the general fit of the model. This is probably because the missing six variables are additional measures of output and prices, that do not offer more information compared to the real GDP and GDP deflator variables that are included in both the eight- and fourteen-variable VARs.

Next, estimation accuracy of the proposed algorithm has to be evaluated. Since the main focus of sign restrictions algorithms is on impulse response analysis, I compare precision of the estimated impulse response functions using all generated datasets. IRFs are combinations of all VAR parameters $\mathbf{\Phi}, \mathbf{\Lambda}, \mathbf{f}, \mathbf{\Sigma}$, therefore comparing their precision provides a convenient summary of overall estimation precision in a VAR model. Fig. 1 shows the responses of the first three variables in the VAR to the three identified pseudo-shocks, in the correctly specified case (C1). Green solid lines are medians over the posterior IRFs in the 500 estimated VARs using an equal number of artificial datasets. Shaded areas show the 90% probability bands of these IRFs. Finally, black dashed lines show the true IRFs implied by the parameters that are fed into the DGP. The 90% bands always include the true IRF, which suggests that estimation precision is satisfactory. The online Appendix shows identical graphs for the four misspecified cases C2–C5. These graphs become a visual confirmation of the numerical results in Table 2, that is, case C5 quite precisely captures the path of the true IRFs, while case C3 results in the largest estimation errors.

3.2. How fast is the new algorithm?

The next Section makes clear that in the context of the empirical application in Furlanetto et al. (2019), the new algorithm is multiple times faster than the algorithm of Rubio-Ramírez et al. (2010) in a six-variable VAR with five identified shocks. Nevertheless, it would be interesting to use artificial data in order to provide more thorough evidence on how fast the factor sign restrictions algorithm is, and how large a VAR it can scale to. For that reason artificial data are generated from the same DGP described in Eqs. (28)–(29) for various values of the key parameters that affect the dimensionality of the VAR, namely T, n and r. Due to the fact that this exercise pushes the VAR dimension n to very large values, I fix $p = 1$ in order to be able to ensure that the VAR process in the DGP is always stationary, and generation of explosive data is excluded. For the purposes of this exercise I set $\mathbf{\Phi} = 0.9\mathbf{I}_n$, $\mathbf{\Lambda}_{ij} \sim U(-1, 1)$ and $\Sigma_i \sim U(0, 1)$, for all $i = 1, \dots, n$ and $j = 1, \dots, r$. During estimation nk sign restrictions are imposed, simply by obtaining the signs of the randomly generated matrix $\mathbf{\Lambda}$.[11]

Table 3 shows the average, over 10 Monte Carlo iterations, machine time in minutes (defined as the total estimation time in seconds divided by 60 and then rounded to the nearest integer) needed to obtain 10,000 draws from the posterior of all parameters after discarding 2000 draws (hence, 12,000 draws in total). These results show that in a huge-dimensional VAR with $n = 100$ series, $T = 500$ observations, and $r = 20$ shocks, it only takes 25 min to obtain 10,000 draws from all parameter matrices, including the 1000 sign-restricted elements in $\mathbf{\Lambda}$. For the smaller model with $n = 15$ – which is already much larger than the vast majority of models considered in the sign restrictions literature – it only takes less than five minutes to obtain the same number of draws when

[10] Information criteria can only be used to compare models with the same dependent variable $\mathbf{y}$.

[11] The purpose of this exercise is not to estimate meaningful restrictions, rather just to measure times. In this case, I impose the maximum number of restrictions possible on $\mathbf{\Lambda}$ in order to test the new algorithm in a worst-case scenario where all nk of its elements are restricted and have to be generated from a truncated Normal posterior.

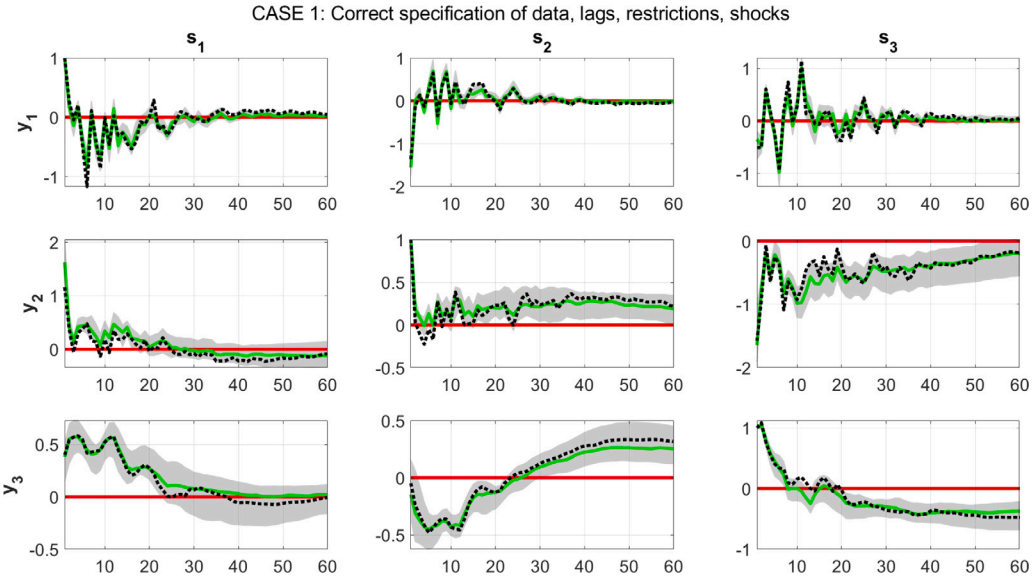


Fig. 1. Impulse response functions of the first three artificially generated variables (denoted as y_1, y_2, y_3) in response to the three identified shocks (denoted as s_1, s_2, s_3) in model C1 (correctly specified model). The green solid lines show the posterior median IRFs over the 500 Monte Carlo iterations, and the gray shaded areas their associated 90% bands. The true IRFs based on the DGP are shown using the black dashed lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Computer time in minutes (defined as (seconds/60), rounded to the nearest integer) for obtaining 10,000 post-burn-in draws (12,000 in total) using various VAR sizes. Here T is the number of observations, n the number of endogenous variables, and r the number of shocks. All VARs have $p = 1$ lag.

	$T = 200$			$T = 500$		
	$n = 15$	$n = 50$	$n = 100$	$n = 15$	$n = 50$	$n = 100$
$r = 3$	1	2	6	4	11	23
$r = 10$	1	4	8	4	12	23
$r = 20$	NA	6	11	NA	15	25

$T = 500$, and only one minute when $T = 200$. These fantastic timings justify the choice to focus on carefully developing a Gibbs sampler that is computationally efficient.[12]

The results above are based on code written in MATLAB2019b and run in a personal computer with Intel Core i7 8700K, tuned at 4.9 GHz, and 32 GB of RAM. Note that the Gibbs sampler algorithm iterates over each VAR equation independently and, thus, significant speed improvements can be achieved by taking advantage of parallel processing abilities of modern computers and high-performance clusters (HPCs). In MATLAB this is as simple as replacing *for loops* with *parfor loops*. Therefore, the algorithm indeed allows the estimation of arbitrarily large VAR models, as it is claimed in the Introduction.

In practical situations, the only issue that might inhibit the performance of the algorithm (and any Monte Carlo-based algorithm, to that effect) is the fact that in very large dimensions we may be sampling parameters $\mathbf{\Phi}$ in a region of the posterior that implies nonstationarity of the VAR. In order to make sense out of impulse response functions, forecast error variance decompositions, historical decompositions etc, we need to make sure we maintain only samples from the posterior which are stationary. For that reason it is important to stress that, throughout my experiments, the horseshoe prior does a great job (especially relative to a subjectively chosen Minnesota prior) in shrinking the coefficients $\mathbf{\Phi}$ towards a more numerically stable region of their posterior, where the VAR model is stationary.

[12] The Gibbs sampler typically loses efficiency when there is high correlation in the samples from the posterior. In the online Appendix I show that, in order to draw $\mathbf{\Lambda}_{ij}$ from univariate (instead of the intractable multivariate) truncated Normal conditionals, we need to condition on $\mathbf{\Lambda}_{-ij}$, i.e. the set of all elements of $\mathbf{\Lambda}$ excluding the ijth. This conditioning increases correlation relative to sampling directly the full matrix $\mathbf{\Lambda}$. However, inefficiency factors for the Gibbs sampler in the linear factor-VAR specification are still quite low (MCMC diagnostic results are available upon request). Additionally, given the ability of the algorithm to obtain quickly tens of thousands of draws from the posterior, concerns about possible correlation of draws can be alleviated by doing "thinning" – i.e. the procedure of storing only every ρth sample from the posterior, where is ρ is the order of the highest significant autocorrelation in the chain.

Table 4
Identified shocks and sign restrictions imposed on the matrix $\mathbf{\Lambda}$ in the baseline, six-variable, five-shock VAR model.

	Identified shocks				
	Supply	Demand	Monetary	Investment	Financial
GDP	+	+	+	+	+
Prices	–	+	+	+	+
Interest rate	NA	+	–	+	+
Investment/output	NA	–	NA	+	+
Stock prices	+	NA	NA	–	+
Spread	NA	NA	NA	NA	NA

Notes: Entries in this table show the restrictions imposed: + positive sign; – negative sign; NA no restriction.

4. Empirics: Financial factors in economic fluctuations

In this section I revisit the empirical exercise in Furlanetto et al. (2019), who aim to measure various financial shocks to the US economy.[13] Given computational restrictions, due to their use of the Rubio-Ramírez et al. (2010) accept/reject algorithm, Furlanetto et al. (2019) end up estimating a series of smaller VARs in order to sequentially measure and label interesting financial shocks, such as uncertainty, credit and housing. Before illustrating how to use the new algorithm to collectively measure all these shocks in one high-dimensional data setting, I first replicate their benchmark results using a smaller VAR. That way, the new algorithm can be contrasted against the output of the Rubio-Ramírez et al. (2010) algorithm, something that is not computationally feasible in the large VAR case.

4.1. Financial shocks using a baseline VAR specification

Among all VAR specifications they use in their work, Furlanetto et al. (2019) specify a *baseline* VAR specification with $p = 5$ lags, using data on real GDP, consumer prices, interest rate, investment-to-output ratio, stock prices, and the external finance premium.[14] All data are for the 1985Q1–2013Q2 period. The online Appendix provides exact details of all series and transformations used, which in this case they are identical to those reported in Table 11 of Furlanetto et al. (2019). Five shocks in total are identified by the authors using the six-variable baseline VAR. The names of the shocks and the associated sign restrictions adopted are shown in Table 4. The first four shocks are standard macro-related shocks, and of interest is the fifth shock which is a generic financial sector shock.

Fig. 2 shows the effects of a financial shock identified as a shock that causes GDP, consumer prices, stock prices, interest rate and the investment/output ratio to react positively contemporaneously. The sign of the spread is left unrestricted. Panel (a) replicates the impulse responses also shown in Figure 1 of Furlanetto et al. (2019), produced using the algorithm of Rubio-Ramírez et al. (2010). Panel (b) shows the same responses produced by application of the new algorithm for sign restrictions. The responses on impact in both panels are of almost identical magnitude, showing that the new algorithm produces sensible results. Any observed differences in the propagation of the impulse responses in subsequent periods, especially for GDP, prices and investment/output ratio, is due to the different estimates of the autoregressive coefficients (Furlanetto et al., 2019 use noninformative priors).[15] Following up on the discussion in the previous section, it takes roughly four hours to obtain 2000 draws from the Furlanetto et al. (2019) using their MATLAB code and exact numerical settings based on the Rubio-Ramírez et al. (2010) algorithm. In contrast, using the same PC[16] it takes 20 min to obtain 600,000 draws from the proposed Gibbs sampler (where out of these 600,000 draws I discard 100,000 and then save every 100th draw, leading to 5000 draws from the posterior of VAR parameters and impulse response functions). Similar conclusions can be made for all other shocks in the system (supply, demand, monetary, investment), where impulse responses are qualitatively similar. Plots for these shocks are provided in the online Appendix.

The new algorithm relies on joint estimation of parameters and identification restrictions. Therefore, it could be argued that the qualitatively similar results in Fig. 2 are an artifact as they can be very sensitive to the structural identification restrictions imposed. However, this is not the case and the algorithm works well in various different scenarios. I consider the following identification restrictions in $\mathbf{\Lambda}$

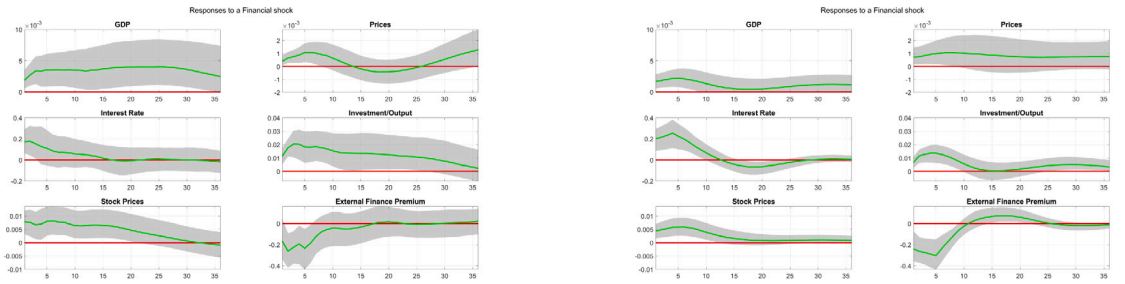
1. Benchmark case with $r = 5$ shocks, identified as in Table 4.

[13] The online Appendix provides the results of an additional numerical exercise (*measuring optimism shocks*) that builds on Arias et al. (2018).

[14] The external finance premium is defined as the spread between yields on Baa rated bonds and the federal funds rate. Notice that the three variables that are not already expressed as rate, ratio, or spread (i.e. GDP, consumer prices, and stock prices), are transformed only using logarithms of the levels and not growth rates. Also note that these authors use a noninformative (uniform) prior, while I use the shrinkage horseshoe prior described in Section 2.

[15] The online Appendix replicates Fig. 2 by exchanging the horseshoe prior for a diffusing (flat) prior. This can be done by simply dropping the local–global hyperparameters of the horseshoe and, instead, setting $\mathbf{V}_i = c \times \mathbf{I}$ for $c \to \infty$. In this case, the normal prior becomes locally uniform on the parameter support. Figure C10 in the online Appendix reveals that in the case of this noninformative prior the shapes of the shocks between the two algorithms become identical. However, the error bands in the new algorithm are still sharper. This is because (Furlanetto et al., 2019) sample the VAR covariance matrix from the standard inverse Wishart posterior (Koop and Korobilis, 2010), while the new algorithm samples the covariance matrix from the more parsimonious factor model.

[16] Specifications of the PC are reported in Section 3.2.



(a) Rubio-Ramírez et al. (2010) algorithm (b) Factor-based sign restrictions algorithm

Fig. 2. This figure replicates the impulse response functions (IRFs) to a financial shock using the baseline specification of Furlanetto et al. (2019). Panel (a) shows results based on the exact configuration of Furlanetto et al. (2019, see Figure 1), using the algorithm of Rubio-Ramírez et al. (2010). Panel (b) replicates the same financial shock using the new sign restrictions algorithm. Solid lines are the posterior medians of the IRFs, and shaded areas the 68% posterior bands.

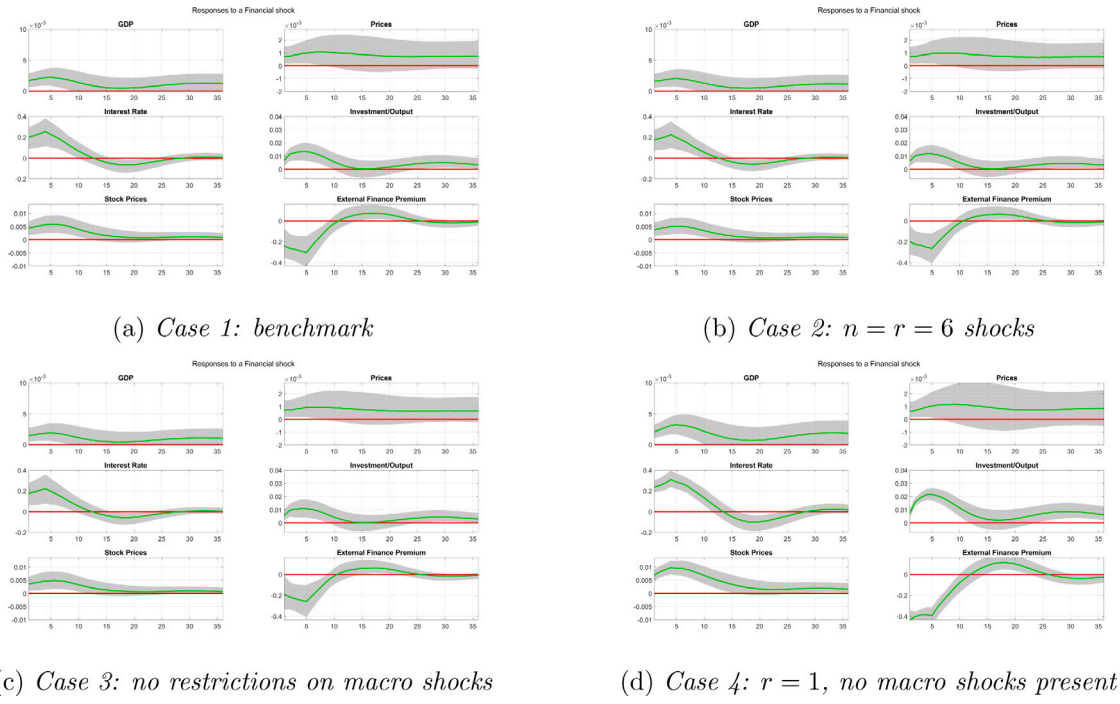
2. Case with $n = r = 6$ shocks. The first five shocks are identical to the benchmark case, and a sixth shock/factor is added with no restrictions. This is equivalent to adding to the restrictions in Table 4 a column with six "NA" entries.
3. Case with $r = 5$ shocks, but only a financial shock is identified. The first four shocks have no restrictions, that is, entries in the first four columns of Table 4 are replaced with NA. The financial shock identified using the restrictions in the last column Table 4.
4. Case with $r = 1$ identified shocks. This is only the financial shock identified using the restrictions in the last column of Table 4, but no macro shocks are identified or estimated.

The first case is the one also plotted in panel (b) of Fig. 2. The second case is used as a means of showing that the algorithm is able to incorporate the case with as many factors as variables, and is not sensitive to the motivating assumption that only a few shocks drive the VAR. This motivating assumption seems reasonable in larger systems (see next subsection), but what if a researcher is interested in smaller systems with as many shocks as variables? The third case allows to find out to what extend identification of the financial shock is affected by the restrictions in the remaining four shocks. Since shocks are identical factors, the aim is to find how estimates of the fifth factor are affected by assumptions in the first four factors. Finally, case four simply removes any information in the first four shocks and simply estimates a model with one shock. Identifying a single shock of interest is a very popular practice in empirical papers that rely on the Rubio-Ramírez et al. (2010) algorithm, as it allows for faster inference (less accept/reject algorithmic steps) and results remain quantitatively unchanged. In contrast, the current algorithm is affected by the assumptions on the number of restrictions. In the fourth case using one shock means that only one factor is used for estimation, which in turn means that estimates of the VAR parameters (covariance matrix, and coefficients of lagged variables) will be affected.

Fig. 3 presents the impulse responses from all four cases. They are qualitatively and quantitatively identical. There are only a couple of differences when moving from the models with five shocks (factors) to the model with only one factor (panel (d) of the figure). In the latter case, the 68% bands of the IRFs of prices are a bit wider, and the median impact response of the spread is twice as large (in absolute value) as the other three cases. Additionally, the curvature of the IRFs of GDP, interest rate, investment/output ratio, and stock prices, over the first 10 periods following the shock, is more pronounced in panel (d) relative to the other three panels. Therefore, Fig. 3 suggests that the number of shocks is more important than identification assumptions made in shocks other than the shock of interest.

In the previous Section, using artificially generated data, it was suggested that it can be hurtful to estimate less shocks compared to the true number of shocks in the DGP. In this case, when estimating $r = 1$ shock does not distort the IRFs substantially, as from a statistical point of view one factor can be sufficient for a small, six-variable VAR. In order to find out if this is truly the case, Table 5 shows the DIC values attained by each of the four cases. The worst case is the one where six shocks are identified from the six series — this corresponds to an overparametrized and unnecessary (from a statistical point of view) factor decomposition. The case with $r = 1$ has the second highest DIC value, meaning that the true number of factors (again in a statistical sense) is larger than one and smaller than six. Surprisingly, Case 3 which is builds on the benchmark Case 1 but lifts all sign restrictions in the first four shocks, is the one that fits the best from a statistical point of view.[17] This DIC value suggests that identifying supply, demand, monetary and investment shocks, using sign restrictions based on economic theory, is statistically inferior to estimating four generic shocks with no sign restrictions. However, the model with the best statistical fit is not necessarily the best model for structural analysis (Bernanke et al., 2005), therefore, in this case the benchmark Case 1 should be preferred as it has a good fit and at the same time allows for identification of important macroeconomic shocks.

[17] Notice that in this case the IRFs of supply, demand, monetary, and investment will be flat around zero — as a matter of fact these should be named as Shock 1, Shock 2, Shock 3, Shock 4 exactly because there are no sign restrictions or identifying assumptions. As it was argued in the previous Sections, even though these four shocks are not structurally identified, the common component $\mathbf{\Lambda f}_t$ is identified and covariance matrix estimation is feasible. Despite the fact that these shocks are not identified, the fifth shock (financial) is identified based on its restrictions and it is not affected by the lack of identification of the first four shocks/factors.



(a) *Case 1: benchmark* (b) *Case 2:* $n = r = 6$ *shocks*

(c) *Case 3: no restrictions on macro shocks* (d) *Case 4:* $r = 1$*, no macro shocks present*

Fig. 3. This figure shows the robustness of the benchmark IRF results produced with the new algorithm, under different identification assumptions. See main text for description of each of the four cases presented in panels (a)–(d). Solid lines are the posterior medians of the IRFs, and shaded areas the 68% posterior bands.

Table 5
Values of the Deviance Information Criterion (DIC) under four different identification schemes (see main text for details). Lower values of the DIC indicate that a given restriction is more plausible than alternatives.

	Case 1	Case 2	Case 3	Case 4
DIC	−8838.02	−7730.67	−9418.08	−8452.36

4.2. A (reasonably) large-scale VAR model for measuring financial shocks

We next proceed to demonstrate how the new algorithm can estimate one, large-dimensional system in order to measure in one setting all the financial shocks that Furlanetto et al. (2019) identify. The larger VAR that these authors specify has seven variables and six shocks: aggregate supply, aggregate demand, investment, housing, uncertainty, and credit. These authors do not identify a monetary shock using this larger VAR, possibly due to computational concerns. Here we attempt to use all available variables in Furlanetto et al. (2019) to identify seven shocks, that is, the six shocks just listed plus a monetary shock. We also use additional measures of output, consumer prices, stock prices, interest rate, and credit spread, in order to enhance identification. We end up with a 15-variable VAR with $p = 5$ on the following variables: (1) real GDP; (2) prices (GDP deflator); (3) interest rate (3-month Tbill); (4) investment-to-output ratio; (5) stock prices (real S&P500 prices); (6) credit spread (Baa minus Fed funds rate); (7) credit to real estate value ratio; (8) excess bond premium (EBP); (9) EBP to VIX ratio; (10) mortgage rate (30-year rates); (11) employment; (12) Federal funds rate; (13) core CPI; (14) stock prices 2 (real DJIA prices); and (15) credit spread 2 ("GZ" spread). The online Appendix has detailed definitions of these variables, transformations used, and sources.

Table 6 shows the signs imposed on each of the 15 variables in order to identify each of the seven structural shocks. This is a large matrix of restrictions, but the new algorithm can handle computationally the task of drawing 600,000 samples from the posterior of all parameters (including the structural matrix of contemporaneous shocks) in a matter of minutes. As it was the case with the baseline VAR above, out of these 600,000 draws 100,000 are discarded and every 100th sample is stored, resulting in 5000 samples used to produce numerical results from this large model. The horseshoe prior also has a crucial role in the estimation of this model, as we have 1140 parameters in $\mathbf{\Phi}$ and only 114 observations for each of the 15 endogenous variables.

Fig. 4 shows the impulse responses of the 15 endogenous variables to a credit shock. The green lines are posterior medians, and the shaded areas 68% bands. The magnitudes and shapes of the IRFs are consistent with the ones reported in Furlanetto et al. (2019, Figure 7), despite the fact that in the case of variables such as GDP the IRFs are strongly different from zero. The most interesting feature of this figure is the effect of a credit shock on the two credit spread variables we used in the same VAR. Furlanetto et al. (2019) use these spreads (plus an additional third spread we have not included here) one at a time in their VAR in order to assess robustness

Table 6
Identified shocks and sign restrictions imposed on the matrix $\mathbf{\Lambda}$ in the large, 15-variable VAR model. These are the restrictions imposed by Furlanetto et al. (2019) using smaller VARs, accumulated into one integrated VAR setting with 15 variables and seven shocks.

	Identified shocks						
	Supply	Demand	Monetary	Investment	Housing	Uncertainty	Credit
Original variables in Furlanetto et al. (2019):							
GDP	+	+	+	+	+	+	+
Prices	–	+	+	+	+	+	+
Interest rate	NA	+	–	+	+	+	+
Investment/output	NA	–	NA	+	+	+	+
Stock prices	+	NA	NA	–	+	+	+
Spread	NA	NA	NA	NA	NA	NA	NA
Credit/real estate value	NA	NA	NA	NA	–	+	+
EBP	NA	NA	NA	NA	NA	–	–
EBP/VIX	NA	NA	NA	NA	NA	+	–
Mortgage rates	NA	NA	NA	NA	NA	NA	–
Additional measures of output, prices etc.:							
Employment	+	+	+	+	+	+	+
Federal funds rate	NA	+	–	+	+	+	+
Core prices	–	+	+	+	+	+	+
Stock prices 2	+	NA	NA	–	+	+	+
Spread 2	NA	NA	NA	NA	NA	NA	NA

Notes: Entries in this table show the restrictions imposed: + positive sign; – negative sign; NA no restriction.

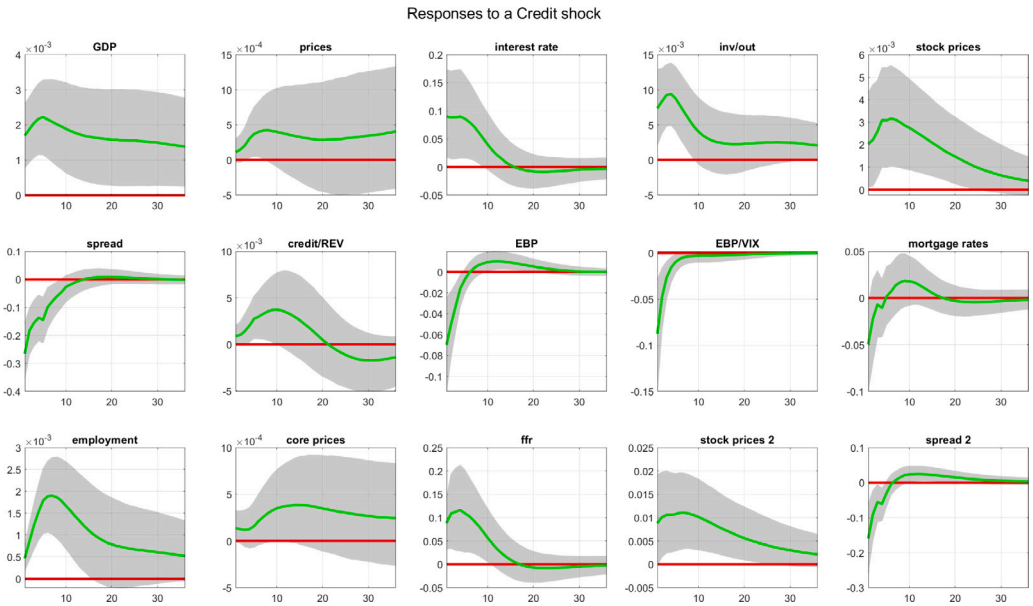


Fig. 4. Impulse response functions to a credit shock in the large, 15-variable VAR with seven shocks identified in total. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of their results. These authors do not impose sign restrictions on the credit spread and they find that in their baseline specification this tends to be negative. In the large VAR case, the first credit spread variable has a strong negative contemporaneous response before subsequently moving to positive territory, while the second credit variable does not have a contemporaneous response different from zero and in subsequent period reacts positively. Such results show the important avenues for identifying various structural shocks that the new algorithm opens up: by using large information sets we can have the ability to identify several structural shocks in one setting, thus, making comparisons and testing of structural hypotheses more transparent. The online Appendix provides further results for this 15-variable VAR.

5. Conclusions

This paper outlines a new algorithm based on a VAR methodology that fully utilizes the interpretability and parsimony of factor models. In particular, the novel element of the proposed approach is the formulation of reduced-form VAR disturbances using a

common factor structure, and the derivation of an algorithm that allows for efficient sampling of sign-restricted decompositions of the VAR covariance matrix. The new algorithm can handle VARs with possibly 100 or more variables and it provides sensible numerical results compared to the algorithm of Rubio-Ramírez et al. (2010) – despite the fact that the two algorithms rely on different modeling assumptions and are not directly comparable.[18] Therefore, the new algorithm can be seen as a useful tool in the toolbox of modern macroeconomists, that complements existing algorithms and at the same time opens up new avenues for empirical research using large-scale VAR models.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.euroecorev.2022.104241.

References

Ahmadi, P.A., Uhlig, H., 2015. Sign Restrictions in Bayesian FaVARs with an Application to Monetary Policy Shocks. Working Paper 21738, National Bureau of Economic Research.
Amir-Ahmadi, P., Drautzburg, T., 2021. Identification and inference with ranking restrictions. Quant. Econ. 12, 1–39.
Anderson, T.W., Rubin, H., 1956. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry. University of California Press, Berkeley, Calif., pp. 111–150.
Arias, J.E., Rubio-Ramírez, J.F., Waggoner, D.F., 2018. Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. Econometrica 86, 685–720.
Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.
Baumeister, C., Hamilton, J.D., 2015. Sign restrictions, structural vector autoregressions, and useful prior information. Econometrica 83, 1963–1999.
Bernanke, B.S., Blinder, A.S., 1992. The federal funds rate and the channels of monetary transmission. Am. Econ. Rev. 82, 901–921.
Bernanke, B.S., Boivin, J., Eliasz, P., 2005. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. Q. J. Econ. 120, 387–422.
Bhattacharya, A., Chakraborty, A., Mallick, B.K., 2016. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. Biometrika 103, 985–991.
Botev, Z.I., 2017. The normal law under linear restrictions: Simulation and estimation via minimax tilting. J. R. Stat. Soc. Ser. B Stat. Methodol. 79, 125–148.
Bruns, M., Piffer, M., 2019. Bayesian Structural VAR Models: A New Approach for Prior Beliefs on Impulse Responses. Tech. rep.
Canova, F., Paustian, M., 2011. Business cycle measurement with some theory. J. Monetary Econ. 58, 345–361.
Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.
Furlanetto, F., Ravazzolo, F., Sarferaz, S., 2019. Identification of financial factors in economic fluctuations. Econ. J. 129, 311–337.
Geweke, J.F., 1996. Bayesian inference for linear models subject to linear inequality constraints. In: Lee, J.C., Johnson, W.O., Zellner, A. (Eds.), Modelling and Prediction Honoring Seymour Geisser. Springer New York, New York, NY, pp. 248–263.
Gorodnichenko, Y., 2005. Reduced-Rank Identification of Structural Shocks in VARs. SSRN working paper 590906, University of California, Berkeley.
Kilian, L., Lütkepohl, H., 2017. Structural Vector Autoregressive Analysis, Themes in Modern Econometrics. Cambridge University Press.
Koop, G., Korobilis, D., 2010. Bayesian multivariate time series methods for empirical macroeconomics. Found. Trends(R) Econom. 3, 267–358.
Matthes, C., Schwartzman, F., 2019. What Do Sectoral Dynamics Tell us About the Origins of Business Cycles?. Working Paper 19-9, Federal Reserve Bank of Richmond.
Ouliaris, S., Pagan, A., 2016. A method for working with sign restrictions in structural equation modelling. Oxford Bull. Econ. Stat. 78, 605–622.
Ramey, V., 2016. Chapter 2 - Macroeconomic shocks and their propagation. In: Taylor, J.B., Uhlig, H. (Eds.), Handbook of Macroeconomics, Vol. 2. Elsevier, pp. 71–162.
Rubio-Ramírez, J.F., Waggoner, D.F., Zha, T., 2010. Structural vector autoregressions: Theory of identification and algorithms for inference. Rev. Econom. Stud. 77, 665–696.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 583–639.
Stock, J.H., Watson, M.W., 2005. Implications of Dynamic Factor Models for VAR Analysis. Working Paper 11467, National Bureau of Economic Research.
Uhlig, H., 2005. What are the effects of monetary policy on output? Results from an agnostic identification procedure. J. Monetary Econ. 52, 381–419.
Uhlig, H., 2017. Shocks, sign restrictions, and identification. In: Honoré, B., Pakes, A., Piazzesi, M., Samuelson, L. (Eds.), Advances in Economics and Econometrics: Eleventh World Congress. In: Econometric Society Monographs, vol. 2, Cambridge University Press, pp. 95–127.

[18] Additional numerical results are provided in the online Appendix.