Cordoni, Francesco; Dorémus, Nicolas; Moneta, Alessio

Working Paper

Identification of vector autoregressive models with nonlinear contemporaneous structure

LEM Working Paper Series, No. 2023/07

Provided in Cooperation with:
Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies

Suggested Citation: Cordoni, Francesco; Dorémus, Nicolas; Moneta, Alessio (2024) : Identification of vector autoregressive models with nonlinear contemporaneous structure, LEM Working Paper Series, No. 2023/07, Scuola Superiore Sant'Anna, Laboratory of Economics and Management (LEM), Pisa

This Version is available at:
https://hdl.handle.net/10419/297115

INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

Identification of Vector Autoregressive Models with Nonlinear Contemporaneous Structure

Francesco Cordoni [1]
Nicolas Doremus [2]
Alessio Moneta [2]

[1] Department of Economics, Royal Holloway University of London, UK
[2] Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy

2023/07 **January 2024**
ISSN(ONLINE) 2284-0400

Identification of Vector Autoregressive Models with Nonlinear Contemporaneous Structure

Francesco CORDONI[1] Nicolas DORÉMUS[2] Alessio MONETA[3]

February 5, 2024

Abstract

We propose a statistical identification procedure for recursive structural vector autoregressive (VAR) models that present a nonlinear dependence (at least) at the contemporaneous level. By applying and adapting results from the literature on causal discovery with continuous additive noise models, we show that, under certain conditions, a large class of structural VAR models is identifiable. We spell out these specific conditions and propose a scheme for the estimation of structural impulse response functions in a nonlinear setting. We assess the performance of this scheme in a simulation experiment. Finally, we apply it in a study on the effects of the macroeconomic shocks that propagate through the economy, allowing for asymmetry between responses from positive and negative impulses.

Keywords: Structural VAR models, Impulse Response Functions, Causal Discovery, Nonlinearity, Additive Noise Models.

JEL classification: C32, C52, E52.

[1]Department of Economics, Royal Holloway University of London. Email: *Francesco.Cordoni@rhul.ac.uk*

[2]Institute of Economics, Scuola Superiore Sant'Anna, Pisa. Email: *nicolas.doremus@santannapisa.it*

[3]Institute of Economics, Scuola Superiore Sant'Anna, Pisa. Email: *a.moneta@santannapisa.it*

The authors acknowledge support from the project "How good is your model? Empirical evaluation and validation of quantitative models in economics" funded by PRIN grant no. 20177FX2A7. We thank Mario Martinoli and Sebastiano Michele Zema for comments on an earlier version of the paper. We are also grateful to the participants of the Italian Workshop of Econometrics and Empirical Economics (Rimini 2022), the Econometric Society European Meeting (Milan 2022), the Workshop "Model Evaluation and Causal Search" (Pisa 2022), and the Italian Congress of Econometrics and Empirical Economics (Cagliari 2023). We also thank Nicoló Maffei Faccioli for providing us with the Excess Bond Premium data series. The R and Matlab code and the data set to replicate the results reported in this article are available at the following repository: https://github.com/f-cordoni/NLAM_time_series.git

1 Introduction

Since the seminal work of Reiersøl (1950), the econometric literature has made clear that linear additive noise models with Gaussian and spherical errors are not identifiable unless a priori restrictions or external instruments are employed. Structural vector autoregressive (VAR) models, widely used in empirical macroeconomic research since Sims (1980), share the same problem of identification. This has been addressed by relying on constraints derived from economic theory (i.e., short-run, long-run, and sign restrictions, as pioneered, respectively, by Sims, 1980, Blanchard and Quah, 1989, and Faust, 1998), by using external instruments (from Romer and Romer, 1989 to Montiel Olea et al., 2021), or, alternatively, by exploiting specific statistical properties of the data that are at odds with the standard model.

This paper contributes to the statistical identification approach. As regards this approach, one finds two strands of the structural VAR literature. One of them exploits non-Gaussianity of the errors, the other exploits heteroskedasticity. In common, they have the use of higher moments for identification (Montiel Olea et al., 2022; Herwartz et al., 2022). This is possible thanks to the introduction of assumptions that depart from the standard model. In this contribution, we want to show that another swerve from the linear model with spherical-normal errors can be exploited for the identification of structural VAR models, namely nonlinearity.

Identification by non-Gaussianity is achieved by the application of ideas and techniques developed in the literature on Independent Component Analysis (ICA), a research field that emerged at the intersection of statistics with signal processing (Comon, 1994; Hyvärinen et al., 2001). Studies in the machine learning literature have shown that ICA can be used for causal inference, and more specifically, for the identification of linear structural equation models, under the assumption of noise independence and non-Gaussianity, as well as the presence of a recursive (i.e. acyclic) structure among the variables (see, e.g., Shimizu et al., 2006, 2011; Hyvärinen, 2013). Moneta et al. (2013) have applied some of these techniques to structural VAR analysis. The structural VAR-ICA literature has been rapidly growing in the recent years, with contributions by Lanne and Lütkepohl (2010); Lanne et al. (2017); Gouriéroux et al. (2017); Herwartz (2018); Fiorentini and Sentana (2023), among others. It is important to note that the recursiveness assumption has been abandoned in these studies so that the impact matrix of the shocks is identified up to the post-multiplication of a generalized permutation matrix, which means, in practice, that one is able to label the shocks only after involving ex post some form of economic reasoning, jointly with an inspection, for instance, of the obtained impulse response functions. Recent studies have also relaxed the independence assumption but maintaining non-Gaussianity and use of higher moments (Lanne and Luoto, 2021; Guay, 2021; Mesters and Zwiernik, 2022). Identification by heteroskedasticity was pioneered by Sentana and Fiorentini (2001) and further pursued by Rigobon (2003); Lanne and Lütkepohl (2008); Sims (2020); Brunnermeier et al. (2021); Lewis (2021), among others.[1]

[1] Identification by heteroskedasticity is a "higher-moment" procedure in the sense of using information from conditional (instead of unconditional) second moments (Montiel Olea et al., 2021), but its underlying model is not, differently from the ICA model, part of the general class of models we consider in this paper.

We claim that nonlinearity, analogously to non-Gaussianity and heteroskedasticity, can be exploited for (recursive) structural VAR identification. This idea has been much less explored in the econometric literature and we believe that it is important to trigger a novel discussion on this. We put forth this idea by — similarly to what has been done in the VAR-ICA approach — importing techniques from the machine learning literature on causal discovery (see the review by Peters et al., 2017) and by presenting a class of autoregressive processes in which it is possible to recover structural shocks (without any problem of labelling) and their dynamic effects on variables of interests directly from the data. In tune with the causal discovery literature, we frame the problem of identifying structural shocks and their contemporaneous impacts in a structural VAR model as a particular case of the general problem of learning a causal structure (which can be usefully represented by a graph) from an observational joint distribution. Hoyer et al. (2008) and Peters et al. (2014) have shown that if the observational distribution follows a structural equation model with an additive noise structure, then the causal graph becomes identifiable from the distribution under specific assumptions. Peters et al. (2014) have also provided practical algorithms to retrieve the causal structure from finite samples.

It is noteworthy that, in the causal discovery framework we introduce, identification by non-Gaussianity and identification by nonlinearity are nested within the general framework of causal discovery with additive noise models. In the first case the key assumptions are, as noted above, shocks' non-Gaussianity plus independence, while in the second case nonlinearity in the shock transmission mechanism plus (again) independence. A related contribution of this paper is to shed light on the connection between these two related identification methods. Furthermore, we spell out the conditions that allow identification where the underlying model is a VAR process with nonlinear dependence at the contemporaneous level. Such a process may or may not show nonlinearities in the autoregressive structure, but the functional autoregressive form is not key for identification. Similarly to the first wave of applications of ICA to VAR analysis (Moneta et al., 2013; Guerini and Moneta, 2017), however, the recursiveness assumption is here required.

We also provide a practical scheme (two algorithms) to recover the structure linking shocks to (reduced-form) VAR innovations and, on the basis of this, to estimate nonlinear structural impulse response functions from time series data. Apart from the role it plays in identification, there are many empirical cases in which the presence of nonlinearity in the processes underlying observed data should not be underestimated. As Ramey (2016) points out, "positive shocks might have different effects from negative shocks, effects might not be proportional to the size of the shock, or the effect of a shock might depend on the state of the economy when the shock hits."

In the next section we present the model set up and the identification methodology, which includes both theoretical results and practical algorithms. In section 3 we present the results of a simulation study. Section 4 presents an empirical application. Section 5 concludes.

2 Identification methodology

In this section, we present our method to achieve statistical identification of a specific class of VAR models, namely VAR models with generic time dependence and nonlinear but recursive contemporaneous causal structure.

In section 2.1, we introduce this class of models, which we call generalized VAR models with additive noise innovations. This class of VAR models is characterised by a recursive causal structure among innovation terms, which can be represented by a directed acyclic graph (DAG). Recovering such causal structure is key for identification, since it allows transforming the reduced-form model into a structural model with independent shocks. In section 2.2, we define and discuss the concept of identification used in this framework.

In section 2.3, we show that the contemporaneous recursive structure among the innovation terms can be recovered by exploiting nonlinearity. In section 2.4, we present an algorithm which infers such structure from estimated reduced-form VAR innovations. In section 2.5, we introduce an algorithm to estimate structural impulse response functions. Theoretical results related to the consistency of the procedure are referred to and discussed in Appendix B.

2.1 Model setup

We consider here a general class of VAR models (see, e.g., the nonlinear structural VAR models in Kilian and Lütkepohl, 2017, ch. 18), in which a vector of K time series variables y_t depends on its lags with an additive vector of disturbances u_t:

$$y_t = F_t(y_{t-1}, \dots, y_{t-L}) + u_t, \tag{1}$$

where $F_t(\cdot)$ is a generic (possibly time dependent) nonlinear or linear function, and u_t is a zero-mean i.i.d. vector of innovations terms and depends only on contemporaneous structural shocks $\varepsilon_{t,1}, \dots, \varepsilon_{t,K}$:

$$u_t = G(\varepsilon_t) \tag{2}$$

We assume that equation (2), for an appropriate ordering of the variables $y_t = (y_{t,1}, \dots, y_{t,K})$, can be written as:

$$u_{t,k} = \varphi_k(\underline{u}_{t,k}) + \varepsilon_{t,k}, \quad \text{for } k = 1, \dots, K \tag{3}$$

where $\underline{u}_{t,k}$ is a subset of the variables $u_{t,1}, \dots, u_{t,k-1}$, and $\varphi_k(\emptyset) = 0$. In other words, u_t can be arranged in a recursive order. We also assume that $\varepsilon_{t,1}, \dots, \varepsilon_{t,K}$ are cross-sectionally mutual independent and that each $\varepsilon_{t,i}$ $(i = 1, \dots, K)$ is i.i.d.. Henceforth, we will refer to the model as formalized in equations (1-3) as the *generalized VAR with additive noise innovation model* (GVAR-ANIM).

Equations (1) - (2) can be seen as a generalization of the standard-linear structural VAR model:

$$y_t = \sum_{\ell=1}^{L} A_\ell y_{t-\ell} + u_t, \tag{4}$$

where

$$u_t = A_0 \varepsilon_t. \tag{5}$$

If $\varepsilon_{t,1}, \ldots, \varepsilon_{t,K}$ are mutually independent and non-Gaussian, equation (5) is an ICA model and the matrix A_0 can be identified up to a post-multiplication of a generalized permutation matrix (Eriksson and Koivunen, 2004; Gouriéroux et al., 2017). If, for any permutation matrix P, the matrix $PA_0^{-1}P^T$ is lower triangular (i.e. A_0^{-1} is *essentially triangular*), under this assumption the matrix A_0^{-1} is uniquely identified (Shimizu et al., 2006)[2]. We call the assumption on A_0^{-1} *recursiveness* assumption, since it imposes a recursive causal structure on the contemporaneous variables of the structural VAR model: if this assumption is true, one can re-order the variables entering in y_t in a "Wold causal chain" (Wold, 1960), so that each variable $y_{t,i}$ causes $y_{t,j}$ and no variable $y_{t,j}$ causes $y_{t,i}$ ($i < j$, for i, j in $1, \ldots, K$).

If A_0^{-1} is essentially triangular, a convenient way to represent the u_t from equation (5) is with a directed acyclic graph (DAG) (see Spirtes et al., 2000; Pearl, 2009). Let us suppose, only for the sake of illustration, that $PA_0^{-1}P^T$ is lower triangular for $P = I$ and $K = 3$. Then we have this system of structural equations between the innovation terms:

$$\begin{cases} u_{t,1} &= \varepsilon_{t,1} \\ u_{t,2} &= \alpha\ u_{t,1} + \varepsilon_{t,2} \\ u_{t,3} &= \beta\ u_{t,1} + \gamma\ u_{t,2} + \varepsilon_{t,3}, \end{cases} \tag{6}$$

which is represented by the DAG in Figure 1.

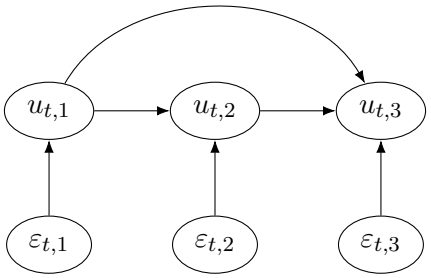

Figure 1: Example of DAG between innovation terms.

If the true DAG were known, one would be able to put (just-identifying or over-identifying) zero restrictions on A_0^{-1} and eventually to recover the independent shocks ε_t (Swanson and Granger, 1997; Demiralp and Hoover, 2003; Moneta, 2008).

Knowing the causal structure is key for identification also in the general nonlinear case of

[2]Shimizu et al. (2006) prove identifiability of linear, non-Gaussian and recursive model (LiNGAM in their terminology) using ICA and, specifically, the Darmois-Skitovich theorem (Comon, 1994, Theorem 11) (see also Peters et al., 2017, Theorem 7.6).

equations (1) - (2). Let us consider the following instance of equation (2), which we have assumed that can be arranged in a recursive order with noises entering in an additive fashion:

$$\begin{cases} u_{t,1} &= \varepsilon_{t,1} \\ u_{t,2} &= \varphi_2(u_{t,1}) + \varepsilon_{t,2} \\ u_{t,3} &= \varphi_3(u_{t,1}, u_{t,2}) + \varepsilon_{t,3} \end{cases} \tag{7}$$

The structure of equation (7) can also be represented by the DAG of Figure 1. Knowing the DAG, or having a method that infers it, allows us to recover the shocks and the manner they impact innovations from observed realizations of u_t. This can be done in a recursive way. Suppose the DAG in Figure 1 is true. Then, following equation (7), one can (i) assign $\varepsilon_{t,1}$ to $u_{t,1}$; (ii) run an adequate (e.g., nonparametric) regression of $u_{t,2}$ on $u_{t,1}$ and get $\varepsilon_{t,2}$ as residual; (iii) regress $u_{t,3}$ on $u_{t,1}$ and $u_{t,2}$ and get $\varepsilon_{t,3}$ as residual. In the next subsections, we will present a method to recover the DAG that represents the causal structure at stake from estimates of u_t in the GVAR-ANIM framework, where nonlinearity (with additive noise) is the general case and linearity (with non-Gaussian noise) is only a special case.

2.2 A note on identification

The concept of identification of linear structural VAR models hinges on the theory of identification in parametric models (Rothenberg, 1971). In this framework, criteria for structural VAR identification have been provided and discussed by Rubio-Ramirez et al. (2010) and Bacchiocchi and Kitagawa (2022). The notion of identification involved in this paper relies on a general formulation, encompassing both parametric and nonparametric models, that has been proposed by Koopmans and Reiersøl (1950) and further developed by Roehrig (1988) (see Matzkin, 2007, for a definition of identification in a purely nonparametric setting). In this general formulation, the target of identification is not (necessarily) a set of parameter points, but, rather, a set of "characteristics" $\mathcal{C}$ of a structure $\mathcal{S} = (f, \Phi)$. A structure consists of a set f of relationships of interest ("structural relationships") between observed variables Y and latent variables V ($f(Y, V) = 0$), and a probability distribution on the latent variables $\Phi(V)$. A structure implies a unique probability distribution Ψ on the observed variable Y. The set of possible structures are restricted a priori by a model $\mathcal{M}$, which is defined as a set of structures which share certain pre-defined characteristics.

Definition (observational equivalence): Two structures in $\mathcal{M}$ are said to be observational equivalent if they imply the same probability distribution Ψ for the observable variables Y (cf. Koopmans and Reiersøl, 1950; Rothenberg, 1971).

Definition (identification): A structure $\mathcal{S}$ is identified in $\mathcal{M}$ if and only if there is no other $\mathcal{S}^* \in \mathcal{M}$ that is observational equivalent to $\mathcal{S}$. A set of characteristics $\mathcal{C}(\mathcal{S})$ is identified in $\mathcal{M}$ if and only if every structure $\mathcal{S}^* \in \mathcal{M}$ that is observational equivalent to $\mathcal{S}$ has the same characteristics $\mathcal{C}(\mathcal{S})$ (cf. Roehrig, 1988, p.435).

Thus, in our case, GVAR-ANIM, as defined in subsection 2.1 limits the set of possible structures to restricted characteristics that will be specified in subsection 2.3. The model is nonparametric in not imposing a parameterization of the functions F_t and G (equations 1 and 2). Therefore, the characteristics that turn out to be identifiable in GVAR-ANIM are not parameter points, but rather the recursive causal structure among the innovations, like the one displayed in equation (7).

2.3 Exploiting nonlinearity

The causal discovery principle that we exploit here was first introduced in the literature by Hoyer et al. (2008). This can be seen as a generalization of the principle underlying causal inference in the linear non-Gaussian case. Let us first show this case, starting from a two-variables model.

Consider a bivariate VAR model, whose vector of reduced-form residuals is $u_t = (u_{t,1}, u_{t,2})$. In the following, to improve readability, we drop the time-subscript t when it is evident from the context and is not relevant, so that, for instance, $u \equiv u_t$ and $u_i \equiv u_{t,i}$. The following result was proven by Peters et al. (2017, Theorem 4.2).

Theorem 1[3] Consider the following linear model:

$$u_1 = \alpha u_2 + \varepsilon_1, \quad \varepsilon_1 \perp\!\!\!\perp u_2. \tag{8}$$

Then there exists $\beta \in \mathbb{R}$ and a random variable ε_2 such that

$$u_2 = \beta u_1 + \varepsilon_2, \quad \varepsilon_2 \perp\!\!\!\perp u_1 \tag{9}$$

if and only if ε_1 and u_2 are Gaussian (where $\perp\!\!\!\perp$ denotes statistical independence).

As a corollary of this theorem, it follows that in a bivariate linear structural VAR, as the one in equation (4) but with $K = 2$ and non-Gaussian shocks, if ε_1 is independent of u_2, then equation (8) is the true structural model and the causal structure can be represented by the DAG $u_2 \longrightarrow u_1$.

Shimizu et al. (2011) extend this result to a multivariate framework. But let us focus here on another possible extension, namely the generalization of this result to the nonlinear case, under the condition that we preserve the noise-additivity assumption. It turns out that, in the strictly nonlinear case, Gaussianity does not preclude identification any longer. Following the terminology by Peters et al. (2017, Definition 4.4), we introduce the additive noise model (ANM) property.

Definition (bivariate ANM): The joint distribution $P(u)$ is said to admit a (bivariate) ANM from u_i to u_j if, for any measurable function φ_j and a variable ε_j we have:

$$u_j = \varphi_j(u_i) + \varepsilon_j, \quad \varepsilon_j \perp\!\!\!\perp u_i \tag{10}$$

Identifiability of a bivariate ANM is based on the following theorem proven by Hoyer et al. (2008, Theorem 1) (see also Peters et al., 2014, 2017):

[3] The proof of this theorem hinges heavily on the Darmois-Skitovich Theorem (see Peters et al., 2017, Theorem 4.3).

Theorem 2 Let us assume that $P(u)$ admits an ANM from u_i to u_j and that ε_j and u_i have strictly positive densities $P(\varepsilon_j)$, and $P(u_i)$, with φ_j, $P(\varepsilon_j)$, and $P(u_i)$ three times differentiable. To simplify notation, let $f := \varphi_j$, and $\xi := \log P(u_i)$ and $\nu := \log P(\varepsilon_j)$, skipping the arguments $u_j - f(u_i)$, u_i and u_i for ν, ξ and f and their derivatives, respectively. Consider the following condition (see Peters et al., 2014, Condition 19).

Condition (C1): the triple $(\varphi_j, P(u_i), P(\varepsilon_j))$ does not satisfy the following differential equation for all u_i, u_j with $\nu''(u_j - f(u_i))f'(u_i) \neq 0$:

$$\xi''' = \xi''(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu'\nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}, \tag{11}$$

If condition C1 is satisfied, then a backward ANM from u_j to u_i, i.e. $u_i = \varphi_i(u_j) + \varepsilon_i$ with $\varepsilon_i \perp\!\!\!\perp u_j$, is not admitted.

Remark. As shown by Hoyer et al. (2008), a "generic" triple $(\varphi_j, P(u_i), P(\varepsilon_j))$ is expected to satisfy condition C1. Zhang and Hyvärinen (2009) actually provide an exhaustive list of five settings that admit an ANM in both the forward and backward direction, the most remarkable case being the linear Gaussian case. Notice that if a joint distribution admits a (bivariate) ANM in one direction (say from u_j to u_i), but not in the backward direction, the DAG representing the ANM is identifiable ($u_j \longrightarrow u_i$).

Figure 2 shows results from an illustrative simulation (see also Peters et al., 2017), in which data x and y are generated from a structural model $y = \beta x + N_y$, with x and N_y following a uniform distribution, with $x \perp\!\!\!\perp N_y$. In the top left panel, we show the results of a linear regression of y on x, whose residuals are plotted over x in the top right panel. It is easy to notice the lack of dependence between such residuals and x. The bottom left panel show the regression with exactly the same data, but performed in the opposite direction, namely we regress x on y. Residuals of such regression are plotted over y (the covariate in the new regression) in the bottom right panel. It is easy to note a (higher-order) statistical dependence. This asymmetry can also be detected in the nonlinear generic case. Figure 3 shows results from another illustrative simulation in which data x and y are generated from a structural model $y = x^3 + N_y$, with x and N_y following a normal distribution, with $x \perp\!\!\!\perp N_y$. In the top left panel, we show the results of kernel regression of y on x, whose residuals are plotted over x in the top right panel. The lack of dependence between such residuals and x is confirmed by a nonparametric independence test, specifically the Hilbert-Schmidt Independence Criterion proposed by Gretton et al. (2007). The p-value for the null hypothesis of independence between residuals and covariate is reported inside the plot. Bottom left panel show the kernel regression with exactly the same data, but performed in the opposite direction, namely we regress x on y. Residuals of such kernel regression are plotted over y (the covariate in the new regression) in the bottom right panel. The p-value for the null hypothesis of independence between residuals and covariate is reported inside the plot and clearly suggests to reject independence.

Peters et al. (2014) proved that the identifiability result stated in Theorem 2 can be extended from the bivariate to the multivariate case. First of all, the definition of ANM given above can be

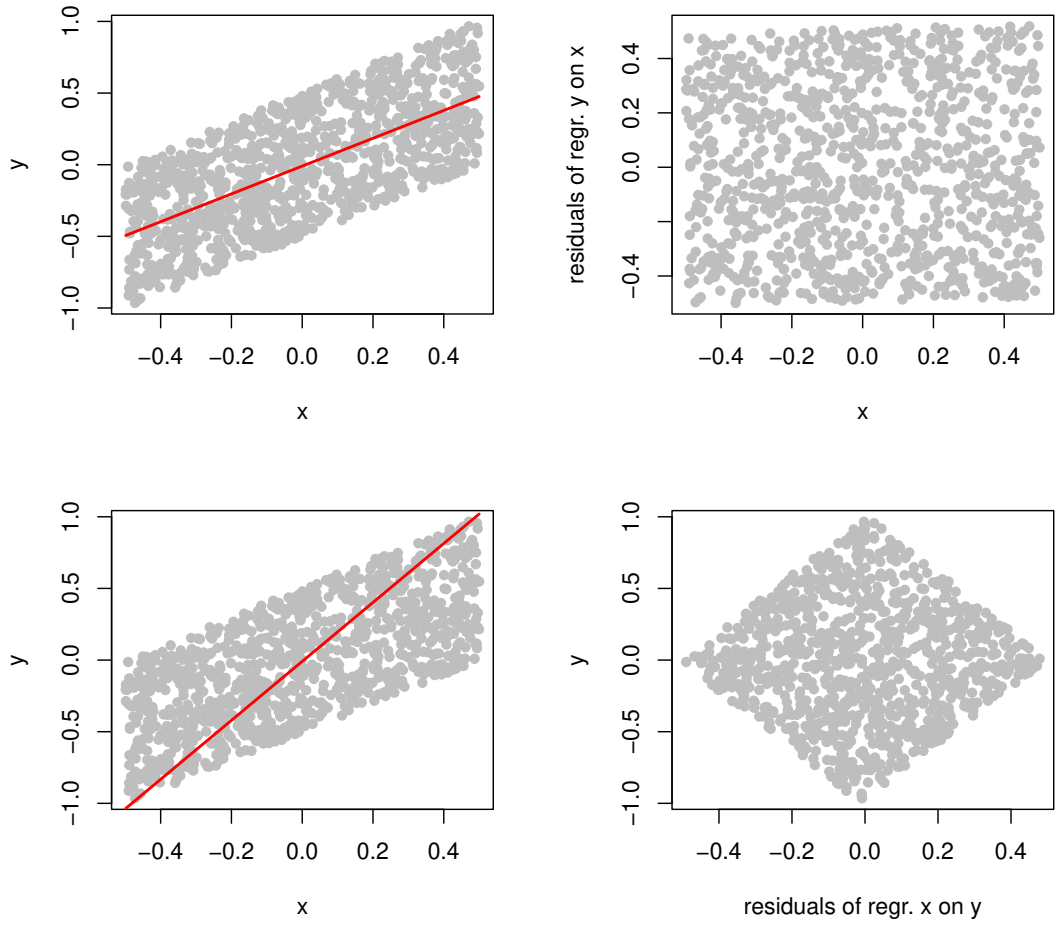

Figure 2: Illustration of the ANM principle with a sample generated from the linear model $y = \beta x + N_y$ with x and N_y independent and drawn from a uniform distribution. Left-side panels: scatter plot of y on x with OLS regression line (red) corresponding to the 'forward' (true) specification (top panel) and to the 'backward' specification (bottom panel). Right-side panels: corresponding regression residuals are plotted with respect to their regressor values, either x (top panel) or y (bottom panel).

straightforwardly extended.

Definition (multivariate ANM). We call a system of K structural recursive equations a (multivariate) ANM if it can be written as:

$$u_k = \varphi_k(Pa(u_k)) + \varepsilon_k \quad \text{for } k \text{ in } 1, \ldots, K, \tag{12}$$

where $\varepsilon_1, \ldots, \varepsilon_K$ are mutually independent. The set of variables $Pa(u_k)$, called *parents* of u_k, is defined as $Pa(u_k) \subseteq \{u_1, \ldots, u_K\} \backslash \{u_k\}$. We denote by $\mathcal{G}$ the DAG representing the structural relations between the u's, i.e. $u_i \longrightarrow u_j$ iff $u_i \in Pa(u_j)$.

Notice that the causal structure $\mathcal{G}$ is a DAG because the system is recursive. Identifiability of

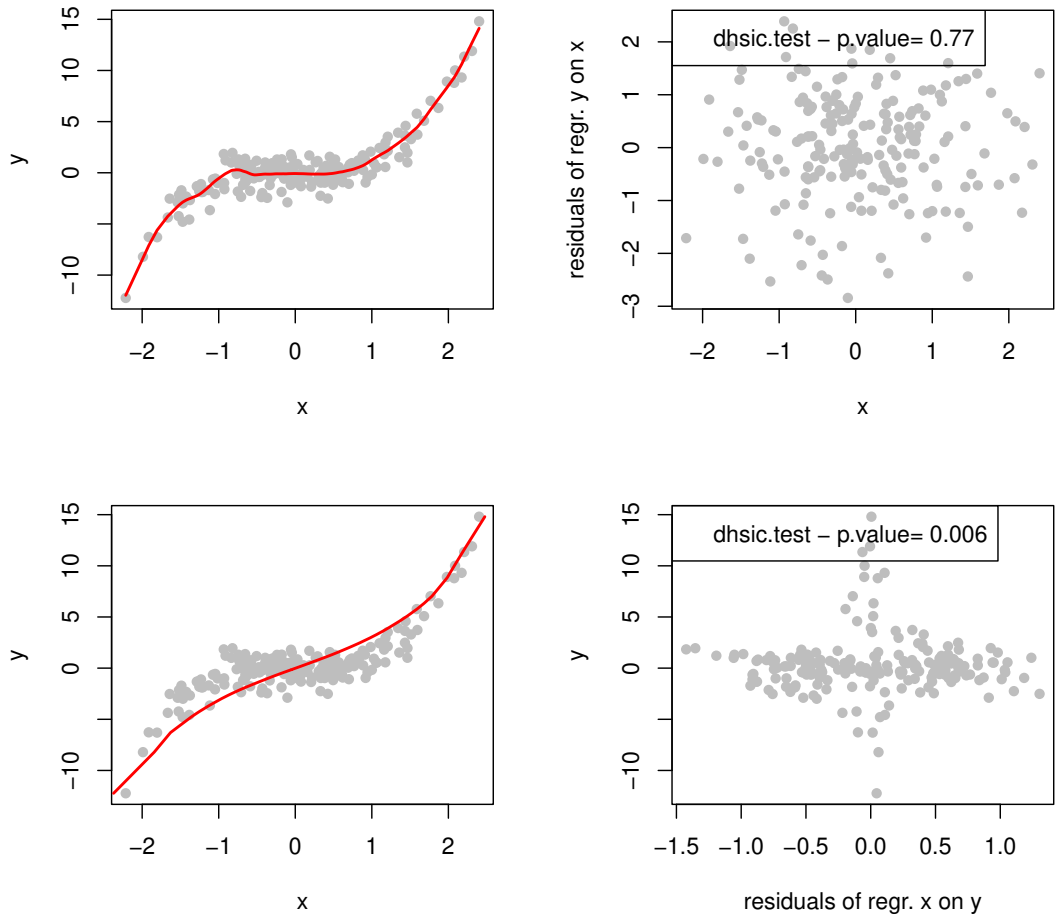

Figure 3: Illustration of the ANM principle with a sample generated from the nonlinear model $y = x^3 + N_y$ with x and N_y independent and drawn from a normal distribution. Left-side panels: scatter plot of y on x with kernel regression line (red) corresponding to the 'forward' (true) specification (top panel) and to the 'backward' specification (bottom panel). Right-side panels: corresponding regression residuals are plotted with respect to their regressor values, either x (top panel) or y (bottom panel).

the multivariate ANM is based on the following theorem:

Theorem 3. Consider a multivariate ANM

$$u_k = \varphi_k(Pa(u_k)) + \varepsilon_k \quad \text{for } k \text{ in } 1, \ldots, K, \tag{13}$$

associated to DAG $\mathcal{G}$ and entailing joint distribution $P(u)$, with φ_k three times differentiable and $P(\varepsilon_k)$ strictly positive for all k. An alternative ANM with DAG $\mathcal{G}'$, where $\mathcal{G}' \neq \mathcal{G}$, also entailing $P(u)$, is not admitted under the following conditions:

Condition (C2) The functions $\varphi_1(\cdot), \ldots, \varphi_K(\cdot)$ are not constant in any of their arguments.

Condition (C3) Let us denote with $ND(u_j)$ the set of graphical "non-descendant" of u_j, i.e.

there is no directed path from u_j to any variable in $ND(u_j)$ (see Spirtes et al., 2000). For all $u_j \in u$, and for all $u_i \in Pa(u_j)$, and all $u_{\mathcal{S}} \in u$ such that $Pa(u_j) \setminus u_i \subseteq u_{\mathcal{S}} \subseteq ND(u_j) \setminus \{u_i, u_j\}$, there is a value $\overline{u_{\mathcal{S}}}$ of $u_{\mathcal{S}}$ with $p(\overline{u_{\mathcal{S}}}) > 0$ such that the triple

$$\left(\varphi_j(\overline{Pa(u_j) \setminus \{u_i\}}, u_i), p(u_i|\overline{u_{\mathcal{S}}}), p(\varepsilon_j)\right) \tag{14}$$

satisfies condition C1. In equation (14) the upper bar indicates a specific value that a random variable or vector takes.

The conditions related to Theorem 3 and its proof can be found in (Peters et al., 2014, see section 3.2 and appendix A therein). Thorem 3 allows us to recover the contemporaneous causal structure among the innovation terms in the GVAR-ANIM. In the next subsection, we will present an algorithm to learn such causal structure from estimated reduced-form VAR residuals.

Topological order

As shown above, knowing the causal structure among innovation terms in the form of a DAG allows specifying the structural model, since the set of graphical parents determines the arguments of $\varphi_k(\cdot)$ in equation (12). Structural shocks can be recovered by regression (parametric or non-parametric) methods applied to the same equation.

Let us now introduce the notion of topological order. Given a DAG over vertices $v_1, \ldots, v_K$, it is possible to associate a total order on its vertices such that if v_i is a parent of v_j (i.e. there is a directed edge from v_i to v_j), then v_i comes before v_j in the ordering. Notice that there is only one topological order associated to a fully connected DAG (a DAG in which each pair of vertices is connected). But, in the case of a sparse DAG, there are possibly multiple topological orders associated to it. Also notice that removing an edge from a fully connected DAG does not invalidate the original topological order, but there could be alternative ordering compatible with it.[4]

Suppose that a set of structural recursive equations, associated to a graph $\mathcal{G}_0$, has generated some data, and suppose that there are multiple topological orders associated to $\mathcal{G}_0$. One can deduce a unique fully connected DAG, which we call $\mathcal{G}^{max}$, for each of these topological orders. It may be the case $\mathcal{G}_0 \neq \mathcal{G}^{max}$. But $\mathcal{G}^{max}$ differs from $\mathcal{G}_0$ only in displaying extra edges, i.e. $\mathcal{G}_0$ is a subgraph of $\mathcal{G}^{max}$. Therefore, if we write down a set of structural equations (as in equation 12) associated to $\mathcal{G}^{max}$, this set of equations will differ from the "true" one, i.e. the one associated to $\mathcal{G}_0$, only in that it includes extra variables in right-hand side. These extra-variables are not in a parent set in (the structural model associated to) $\mathcal{G}_0$. But since the original causal ordering (associated to $\mathcal{G}_0$) is maintained, no endogeneity issues arise when one uses $\mathcal{G}^{max}$ as a template for a regression model. Indeed, if one applies regression methods to the set of equations associated to $\mathcal{G}^{max}$, which may therefore include extra variables (with respect to the "true" structural model) on

[4]Consider the fully connected DAG on v_1, v_2, v_3 associated to the topological order $\langle v_1, v_2, v_3 \rangle$. From removing $v_1 \longrightarrow v_3$, we get $v_1 \longrightarrow v_2 \longrightarrow v_3$, and the original topological order $\langle v_1, v_2, v_3 \rangle$ is still the only one valid. From removing $v_2 \longrightarrow v_3$, we get $v_2 \longleftarrow v_1 \longrightarrow v_3$, and there are two possible topological orders: $\langle v_1, v_2, v_3 \rangle$, $\langle v_1, v_3, v_2 \rangle$.

the right hand side, one obtains consistent estimates, since these variables are at most irrelevant, but do not entail reverse causality problems.

In practice, if the aim of the study is to estimate the structural shocks (e.g., in impulse response function analysis), one may aim at recovering a topological order instead of a DAG, so that to minimize inference errors in finite samples. Instead, if causal discovery is the goal, then a topological order is not enough. This is why in the next section we will present a search algorithm that can be stopped at the topological order step.

A simple case

While the paper, including the algorithms presented in the next subsections, covers a wide class of models that we have labelled GVAR-ANIM (see equations 1-3), in both the simulation and the empirical analysis, we are going to focus on a simpler case, namely the case of linear lagged relationships with nonlinear contemporaneous relationships between innovations (with additive structural shocks). In terms of equations:

$$\begin{cases} y_t &= \sum_{\ell=1}^{L} A_\ell y_{t-\ell} + u_t, \\ u_{t,k} &= \varphi_k(Pa(u_{t,k})) + \varepsilon_{t,k} \quad \text{for } k \text{ in } 1, \ldots, K. \end{cases} \tag{15}$$

We decide to focus on this simple case because the crucial issue in VAR identification (see, e.g. Stock and Watson, 2001) is the solution of the contemporaneous causality problem, which allows to estimate the structural shocks. This holds both in the linear and nonlinear VAR case. Since the method we propose exploits contemporaneous nonlinearity, equation (15) represents the most simple time series model in which we can apply it.

2.4 Algorithm for the contemporaneous structure

Previous sections showed that, under the assumption of additive noise, the (graphical) causal structure of a recursive structural equation model is generically identifiable. There are alternative strategies that exploit this identifying principle using sample data (see Peters et al., 2014 and Peters et al., 2017 for an overview). Here we are presenting the RESIT (REgression with Subsequent Independence Test) algorithm proposed in Peters et al. (2014), which incorporates the principles of identification in the nonlinear additive noise setting directly and straightforwardly. It also has the advantage of segmenting the causal search in, first, a topological order search phase and, subsequently, a DAG search phase. This modular architecture fits well with the possibility that for shock recovery (and impulse response analysis) a topological order can be sufficient for the scope (in finite samples), as noted above. The original RESIT algorithm of Peters et al., 2014 operates in two phases (see Appendix A). We add a preliminary phase, which we call Phase 0, in which we estimate a reduced-form GVAR-ANIM (see equation 1) and use its estimated residuals as input for the next phase.

Phase 1 aims at ranking the input variables in a topological order. The topological order is elicited in a backward fashion by determining which variable is the latest (i.e. a sink node) among the whole set of variables to be ordered, removing it from the set and repeating the search over the resulting subset until only one element remains, which is given the first position in the ranking. At each iteration, the search is performed by regressing each variable jointly on all the others and measuring the dependence (more on this below) between the regressors and the obtained residual. The regressand yielding the weakest dependence is regarded as the sink node among the current set of candidates; it is removed from the set of candidates and is assigned the first position among the set of variables that have already been removed. The topological order simply corresponds to the reversed order of elimination. The topological order determines, for each variable, a set of potential parents. This set coincides with all the variables ranked higher.

Phase 2 (which we call *pruning phase*) further narrows down the sets of potential parents by performing variable selection (see Appendix A). In other words, it prunes edges from the fully connected DAG which one can draw from the topological order of Phase 1. It essentially performs a variable selection task. It regresses each variable on its putative parents (according to the topological order derived above) but omitting iteratively one of them. If the estimated residual of the corresponding nested model is found independent of all preceding variables in the topological order, the omitted regressor is removed from the set of parents of the regressand. The search stops when no regressor can be removed without creating dependence.

Any regression method and dependence measure can be embedded in the RESIT algorithm, as long as they are adequate with respect to the data-generating process. Not knowing the functional dependence, i.e. $\varphi_k(\cdot)$ in equation (12), an appropriate strategy would be to use a nonparametric kernel estimator (Fan and Gijbels, 2018). In our implementation, in tune with Hoyer et al. (2008), we opt for the Gaussian process approach (see Williams and Rasmussen, 2006). An alternative, more restricted, nonparametric model is the generalized additive noise model (Hastie and Tibshirani, 2017). At the extreme, the functional form can be "restricted" down to linearity, as long as one assumes non-Gaussianity of the noises. In this case, the ANM collapses to the Linear Non-Gaussian Acyclic Model studied by Shimizu et al. (2006) (see also Shimizu et al., 2011).

As regards the dependence measure, any measure that accounts for higher order statistics (rather than linear dependence only) can in principle work. This feature is essential because regression techniques that minimize quadratic errors including linear components yield estimated noises that are orthogonal to the covariates. Peters et al. (2014) propose to use the p-value of a nonparametric independence test as dependence score, so that the dependence is minimized when the p-value, under the null hypothesis of independence, is maximized. This follows the principle of Hodges-Lehmann estimation, which was proposed by Herwartz (2018) for independent component analysis. The specific independence test that we perform in our implementation, following (Peters et al., 2014), is the kernel independence test proposed by Gretton et al. (2007) with the Hilbert-Schmidt Independence Criterion (HSIC). The asymptotic results presented by these authors guarantee a convergence of the p-value to zero only under dependence.

We mentioned above that the RESIT algorithm can be stopped at the topological order step

(i.e. Phase 1) if the aim is to estimate structural shocks and impulse response functions. The second phase should instead be applied if one is interested in uncovering the exact causal structure among the innovation terms. Since both in our simulation study and in the empirical application we focus on structural impulse response analysis, the algorithm shown in the main body of the text skips the edge pruning phase (see Algorithm 1 here below). In Appendix A, one can find the original RESIT algorithm embedded in a more general procedure that deals with time series data. Notice that the motivation to skip Phase 2 of RESIT is only of practical nature, when one deals with finite sample. Contrary to Phase 1, Phase 2 operates on the basis of statistical hypothesis testing (independence test), in which one has to decide a rejection threshold. In finite sample, this decision may lead to mistakenly classify weak dependence as independence and therefore to mistakenly cutting edges between directly causally related variables. For the sake of retrieving structural shocks from reduced-form VAR residuals, one has to run a battery of regressions of the form expressed in equation (13), in which shocks will be the residuals of these regressions. If we omit any variable from the parents' set (i.e. covariates of these regressions) because we have made a mistake in Phase 2, we introduce an omitted variable bias. But if we keep an irrelevant variable in the parents' set, we may add some variance in the estimates but we do not introduce any bias. Given this trade-off, we opt for not using Phase 2 for the sake of estimating impulse response functions.

Algorithm 1 VAR + RESIT (Phase 1 is from Peters et al., 2014)

Input: A K-dimensional time series vector $(y_{t,1}, \ldots, y_{t,K})'$
PHASE 0: Estimate the reduced-form model.
Estimate a reduced-form time series model of the class GVAR-ANIM, see eq. (1) and extract residuals $\widehat{u}_t = (\widehat{u}_{t,1}, \ldots, \widehat{u}_{t,K})'$. To simplify the notation, let us call $\widehat{u}_{t,k} \equiv v_k$, for $k = 1, \ldots, K$.
$S := 1, ..., K$, $\pi := [\quad]$

PHASE 1: Determine topological order and potential parent set
repeat
    for $k \in S$ **do**
        Regress v_k on $\{v_i\}_{i \in S \setminus \{k\}}$ and obtain residuals e_k
        Measure dependence between e_k and $\{v_i\}_{i \in S \setminus \{k\}}$
    end for
    Let k^* be the k with the weakest dependence
    $S := S \setminus k^*$
    $pa(k^*) := S$
    $\pi := [k^*, \pi]$
until $\#S = 0$
Output: π
Output: $(pa(1), \ldots, pa(K))$

Note: with $pa(k)$ we refer to the set of indices associated to the variables in $Pa(v_k)$, for any k.

A simple illustration

Let us suppose that a 3-variable GVAR-ANIM process of the type of equations (1-3) has generated some data. Specifically, suppose that equation (3) corresponds to:

$$\begin{cases} u_{t,1} = \varepsilon_{t,1} \\ u_{t,2} = \varphi_2(u_{t,1}, u_{t,3}) + \varepsilon_{t,2} \\ u_{t,3} = \varepsilon_{t,3} \end{cases} \tag{16}$$

The DAG associated to these equations is $u_{t,1} \longrightarrow u_{t,2} \longleftarrow u_{t,3}$. Now, consider to have estimated $u_{t,1}, u_{t,2}, u_{t,3}$ as reduced-form residuals of a VAR of the form of equation (1), following Phase 0 of Algorithm 1. A mistake-free output of Phase 1 will be either the topological order $\langle u_{t,1}, u_{t,3}, u_{t,2} \rangle$ or $\langle u_{t,3}, u_{t,1}, u_{t,2} \rangle$. A mistake-free output of Phase 2 (see Appendix A) will be $u_{t,1} \longrightarrow u_{t,2} \longleftarrow u_{t,3}$. In order to estimate the structural shocks $\varepsilon_1, \varepsilon_2, \varepsilon_3$ one can run a set of regressions on the basis of equations (16). But suppose one stops at Phase 1. Then one would estimate one of the following regression models instead:

$$\begin{cases} u_{t,1} = \varepsilon_{t,1} \\ u_{t,2} = \varphi_2(u_{t,1}, u_{t,3}) + \varepsilon_{t,2} \\ u_{t,3} = \varphi_3(u_{t,1}) + \varepsilon_{t,3} \end{cases} \tag{17}$$

$$\begin{cases} u_{t,1} = \varphi_1(u_{t,3}) + \varepsilon_{t,1} \\ u_{t,2} = \varphi_2(u_{t,1}, u_{t,3}) + \varepsilon_{t,2} \\ u_{t,3} = \varepsilon_{t,3} \end{cases} \tag{18}$$

Both systems of equations (17) and (18) differ from (16) in including irrelevant variables in the regression systems, but not involving any reverse causality issue.

2.5 Algorithm for the structural impulse response functions

The advantage of identifying a structural model is that one can estimate the dynamic effects of structural shocks. But, under the generic GVAR-ANIM framework, this cannot be done by using the conventional tools of impulse response analysis. Indeed in such a framework the variables' responses to a specific shock at a certain time depend on the system's past history, sign and magnitude of all contemporaneous shocks and subsequent ones. Following Koop et al. (1996) and Kilian and Lütkepohl (2017, ch.18), we define the *structural* (nonlinear) impulse response function (IRF) as:

$$\text{IRF}(h, \delta, \Omega_{t-1}) = E(y_{t+h} | \varepsilon_{t,i} = \delta, \Omega_{t-1}) - E(y_{t+h} | \Omega_{t-1}), \tag{19}$$

where δ is the (positive or negative) magnitude of the shock $\varepsilon_{t,i}$, whose effects one wants to study, Ω_{t-1} is the history of the model data up to time $t-1$, and h is the horizon point up to which the impulse response function is studied. To estimate equation (19), we can use the Monte Carlo integration approach suggested by Kilian and Lütkepohl (2017).

The idea of this approach, whose detailed scheme is presented in Algorithm 2, is to take pairs of simulation runs: $\tilde{Y}^{\delta,n}$ and $\tilde{Y}^{n}$ (for $n = 1, \ldots, N$), in which $\tilde{Y}^{\delta,n}$ denotes the "treated" time series and $\tilde{Y}^{n}$ the "control" one, where N is the number of Monte Carlo runs. Both elements share a common history Ω_{t-1}, but diverge at date t, in which a structural shock of interest is set to δ in one run (treatment), while no such restriction applies to the other (control). Apart from this restriction, all shocks are free to fluctuate in both groups of series from time t onwards. The difference is then taken for each pair. Finally, estimated impulse response functions are calculated by taking average of these simulated differences.

Algorithm 2 requires as input estimates (from the observed data) of the functional relationship $\widehat{\varphi}_k(\cdot)$ between innovations in $u_{t,k} = \varphi_k(Pa(u_{t,k})) + \varepsilon_{t,k}$ (for k in $1, \ldots, K$); estimates of the lagged functional relationship $\widehat{f}_k(\cdot)$ in $y_{k,t} = f_k(y_{t-1}, \ldots, y_{t-p}) + u_{k,t}$ (for k in $1, \ldots, K$);[5] estimates of the structural shocks $\varepsilon_{t,1}, \ldots, \varepsilon_{t,K}$. These can be estimated through the following procedure. For each $k = 1, \ldots, K$, if $Pa(u_{t,k}) = \emptyset$ then set $\widehat{\varepsilon}_{t,k} := \widehat{u}_{t,k}$; else (non-parametrically) regress $\widehat{u}_{t,k}$ on $\{\widehat{u}_{t,i}\}_{i \in pa(k)}$ and set $\widehat{\varepsilon}_{t,k} := \widehat{u}_{t,k} - \widehat{\varphi}_k(\{\widehat{u}_{t,i}\}_{i \in pa(k)})$.[6]

Each simulated[7] path ($\tilde{Y}^{\delta,n}$ and $\tilde{Y}^{n}$) is computed iteratively at each time period, starting from the time point of the "treatment". At each step, new values are generated summing (i) lagged effects $f_k(y_{t-1}, \ldots, y_{t-L})$; with (ii) innovation terms $u_{t,k} = \varphi_k(Pa(u_{t,k})) + \varepsilon_{t,k}$, for every k. New values of $\varepsilon_{t,k}$ are independently drawn from their marginal empirical distribution at each time horizon (except for the one set at δ at time t in the treatment run). Since the innovation model is recursive, each $\tilde{u}_{t,k}$ is computed iteratively at each time step by adding its structural shock $\varepsilon_{t,k}$ to the contemporaneous effect of its parents ($Pa(u_{t,k})$). Parents' sets can be assigned in two alternative ways: (i) by simply taking the parents' sets as given by the full version of RESIT as it appears in Appendix A; or (ii) in a more conservative fashion (see discussion above), by considering the sole topological order (see Algorithm 1 in section 2.4) and by regarding any variable $u_{t,k}$ as a parent of those it precedes in that ordering. Either way, the functional forms $\widehat{\varphi}_k(\cdot)$ of the contemporaneous effects are estimated via (nonparametric) regression of each variable over the parent set considered.

3 Simulation analysis

The simulation study documented in this section aims at evaluating the performance of our approach in estimating structural impulse response functions. We simulate a simple VAR with different contemporaneous causal structures and different parametrizations. We then estimate the structural IRFs using Algorithm 2, but under different identification approaches, and, we systematically compare them. Our simulation results, as reported below, demonstrate that, in case of nonlinearity in the data generating process, our procedure is able to satisfactorily estimate the

[5] For simplicity, we dropped the time index from the lagged effects $f_{k,t}(\cdot)$. A time-varying function can be incorporated into the algorithm we are discussing in this subsection by updating the $f_{k,t}(\cdot)$ at each step.

[6] With $pa(k)$ we refer to the set of indices associated to the variables in $Pa(u_{t,k})$, for any k.

[7] Simulated variables are marked with a tilde sign ($\sim$).

Algorithm 2 Structural IRFs at time t^*

1: **Setting Values**:

- t^*: date of the shock to be studied
- H: max horizon
- k^*: index of the shocked variable
- δ: shock magnitude
- N: number of Monte Carlo runs

2: **Input data**

- history (Ω_{t^*-1}) at date t^*
- estimated structural shock matrix $\widehat{E}$
- estimated $\widehat{f_k}(\cdot)$, $\widehat{\varphi}_k(\cdot)$ (for each k)
- topological order: $\pi = (\pi_1, \ldots, \pi_K)$; parents set: $(pa(1), \ldots, pa(K))$

3: PHASE 1: Simulate N couples of time paths and get the difference
4: **for** $n \in \{1, \ldots, N\}$ **do**
5: Create a $(t^* + H) \times K$ empty matrix $\tilde{Y}^n$. Set $1 : (t^* - 1)$ rows $:= \Omega_{t^*-1}$
6: $\tilde{Y}^{\delta,n} := \tilde{Y}^n$
7: Create a $(H + 1) \times K$ matrix $\tilde{E}$, randomly sampling from $\widehat{E}$ (columnwise)

Note: denote any (t, k) entry of any matrix E with $E_{t,k}$

8: Create a $(H + 1) \times K$ matrix $\tilde{E}^\delta$, as in line 7, but set $\tilde{E}^\delta_{t^*,k^*} := \delta$
9: Create a $(H + 1) \times K$ matrices $\tilde{U}$ and $\tilde{U}^\delta$(initially empty)
10: Create a $(H + 1) \times K$ matrices $\tilde{I}^{n;k^*,\delta}$ (initially empty)
11: **for** $h \in \{0, \ldots, H\}$ **do**
12: **for** $k \in \{\pi_1 \ldots \pi_K\}$ **do**
13: $\tilde{U}^\delta_{h,k} := \widehat{\varphi_k}(\{\tilde{U}^\delta_{h,i}\}_{i \in pa(k)}) + \tilde{E}^\delta_{h,k}$ (with $\varphi_k(\emptyset) = 0$)
14: $\tilde{Y}^{\delta,n}_{t^*+h,k} := \widehat{f}_k(\tilde{Y}^{\delta,n}_{t^*+h-1}, \ldots, \tilde{Y}^{\delta,n}_{t^*+h-L}) + \tilde{U}^\delta_{h,k}$
15: $\tilde{U}_{h,k} := \widehat{\varphi_k}(\{\tilde{U}_{h,i}\}_{i \in pa(k)}) + \tilde{E}_{h,k}$
16: $\tilde{Y}^{n}_{t^*+h,k} := \widehat{f}_k(\tilde{Y}^{n}_{t^*+h-1}, \ldots, \tilde{Y}^{n}_{t^*+h-L}) + \tilde{U}_{h,k}$
17: $\tilde{I}^{n,k^*,\delta}_{h,k} := \tilde{Y}^{\delta,n}_{t^*+h,k} - \tilde{Y}^{n}_{t^*+h,k}$
18: **end for**
19: **end for**
20: **end for**
21: PHASE 2: Average values
22: $\tilde{I}^{k^*,\delta,t^*} := \frac{1}{N} \sum_{n=1}^{N} \tilde{I}^{n,k^*,\delta,t^*}$
23: **Output**: $\tilde{I}^{k^*,\delta,t^*}$

structural IRFs and that a linear approximation may lead astray even under correctness of the recursive order.

3.1 Data generating processes

We simulate a 3-variables VAR model with one lag, which can be written as

$$y_t = A_1 y_{t-1} + u_t, \tag{20}$$

where A_1, to control for the persistence of the process, is a lower triangular matrix with all elements below and on the main diagonal equal to 0.5. The u_t terms are generated following three types of causal structures, which we call the "causal chain" (21), "common cause" (22), and "v-structure" (23):

Causal chain

$$\begin{cases} u_{t,1} = \varepsilon_{t,1} \\ u_{t,2} = \varphi_2(u_{t,1}) + \varepsilon_{t,2} \\ u_{t,3} = \varphi_3(u_{t,2}) + \varepsilon_{t,3} \end{cases} \tag{21}$$

Common cause

$$\begin{cases} u_{t,1} = \varepsilon_{t,1} \\ u_{t,2} = \varphi_2(u_{t,1}) + \varepsilon_{t,2} \\ u_{t,3} = \varphi_3(u_{t,1}) + \varepsilon_{t,3} \end{cases} \tag{22}$$

v-structure

$$\begin{cases} u_{t,1} = \varepsilon_{t,1} \\ u_{t,2} = \varepsilon_{t,2} \\ u_{t,3} = \varphi_3(u_{t,1}, u_{t,2}) + \varepsilon_{t,3} \end{cases} \tag{23}$$

Furthermore, each generic model will be given three alternative parametrizations to $(\varphi_\bullet(\cdot), \varepsilon_\bullet)$. We first define a "linear Gaussian" setting in which each $\varphi_\bullet(x) = x$ and $\varepsilon_{t,1}, \varepsilon_{t,2}, \varepsilon_{t,3}$ are mutually independent and (standard) normally distributed. Second, the "linear non-Gaussian" setting is characterized by the same identity function but the $\varepsilon_{t,1}, \varepsilon_{t,2}, \varepsilon_{t,3}$ are drawn from a Laplace distribution with scales parameters $1, 2, 4$, respectively. Finally, we define a "nonlinear Gaussian" setting in which shocks are drawn in the same fashion as in the "linear Gaussian" setting and the functional forms taken by $\varphi_2(\cdot)$ and $\varphi_3(\cdot)$ in each causal structure are reported in Table 1. We generate samples $\{y_t\}_{t=1}^T$ of size $T = 250, 500, 1000$.

Table 1: Functional forms of nonlinear models. The parameters α and β are independently drawn from a uniform distribution with support $[1, 4]$.

	Causal chain	Common cause	v-structure
$\varphi_2()$	$\text{sign}(u_{t,1})\lvert u_{t,1}\rvert^\alpha$	$\text{sign}(u_{t,1})\lvert u_{t,1}\rvert^\alpha$	$\text{sign}(u_{t,1})\lvert u_{t,1}\rvert^\alpha$
$\varphi_3()$	$\sin(\text{sign}(u_{t,2})\lvert u_{t,2}\rvert^\beta)$	$\sin(\text{sign}(u_{t,2})\lvert u_{t,2}\rvert^\beta)$	$\text{sign}(u_{t,1})\lvert u_{t,1}\rvert^\alpha + \sin(\text{sign}(u_{t,2})\lvert u_{t,2}\rvert^\beta)$

3.2 Identification schemes

For each selected T, type of structure and parametrization we generate 200 artificial (3-variate) time series, from which we estimate IRFs. These are constructed following Algorithm 2, according to five different schemes, which we call: (i) "theoretical" IRFs, (ii) Cholesky-based (linear) IRFs, (iii) true-graph-based nonlinear IRFs, (iv) true-topological-order-based nonlinear IRFs, and (v) Algorithm-1-based nonlinear IRFs.

Scheme (i) builds IRFs by applying Algorithm 2 to each artificial time series, except that instead of the estimated structural shock matrix (see "Input data" in the pseudo code) we use the true DGP's one, instead of the estimated functional forms we use the autoregressive matrix A_1, and the true generating functions $\varphi_k(\cdot)$, and we use the correct set of parents (and topological order) implied by the true causal structure. We calculate IRFs using this scheme because we want to have a ground truth against which we can assess our procedure. In practice, scheme (i) is the application of Algorithm 2 to the true model, bypassing the model estimation step. For this reason, we call the produced IRFs "theoretical".

In contrast, the remaining four procedures are based on estimates of the model. Scheme (ii) is meant to provide a benchmark to the approaches that, differently from this scheme, allow for nonlinearity. Here structural shocks are computed via Cholesky factorization of the reduced-form residuals' covariance matrix, where the recursive ordering of the innovation structure is the true one, instead of being determined in a data driven fashion as Algorithm 1 does. Moreover, for the sake of comparison with the other schemes, we do not apply here the standard linear approach that derives IRFs from MA coefficients, but we feed Algorithm 2 with the estimated linear coefficients. We label IRFs from this scheme "CHOL" (from Cholesky).

Scheme (iii) is similar to scheme (i) except that the functional (autoregressive and contemporaneous) dependencies are estimated from the data, on the basis of the true causal graph (which automatically implies a correct topological order). Functional forms are estimated via OLS as regards the autoregressive components and through Gaussian process regression as regards the contemporaneous structure. In practice, in this scheme we are bypassing the data-driven causal (order) search by using the correct specification of the model. The aim here is to isolate the performance of Algorithm 2, avoiding potentially compounding mistakes of Algorithm 1. We label IRFs from this scheme "TDAG" (i.e. true DAG).

Scheme (iv) is a slight variant of scheme (iii): we feed Algorithm 2 with a correct topological order and a fully connected DAG deduced from it, which may have extra edges with respect to the true DAG. Thus, we also isolate here the performance of Algorithm 2 (as in (iii)), but we allow the introduction of potentially irrelevant variables in its inputs, to see if they make a difference in terms of performance. Notice that Algorithm 1, which is involved in the next scheme, may also introduce these extra variables, so that scheme (iv) is a closer benchmark to scheme (v). Furthermore, scheme (iv) is a nonlinear counterpart of the Cholesky scheme (ii) because they are both based on the true topological order, without any further restrictions (i.e., they may include irrelevant covariates). We label IRFs from this scheme "TTOP" (i.e. true topological order).

Scheme (v) implements the data-driven approach proposed in Section 2 so that its assessment is the focus of this analysis, the previous schemes ultimately being benchmarks for this one. Here, from each generated time-series sample we estimate the topological order, from which we deduce a fully connected DAG and a corresponding set of parents, by Algorithm 1 and subsequently we apply Algorithm 2. As input for the functional forms we use OLS for the autoregressive part (as in schemes (ii-iii-iv)) and Gaussian process regression for the contemporaneous structure (as in schemes (iii-iv)). Recall from Section 2.5 that the structural shocks (further input of Algorithm 2)

are recovered from the residuals of the regression model at the contemporaneous level. We label IRFs from this scheme "ALG1" (i.e. based on Algorithm 1).

3.3 Simulation results

Let us call $\text{IRF}^{\bullet}_{k,h}$ the IRF referring to the responses of variable $y_{t,k}$ to shock $\varepsilon_{t,1}$, at horizon h, using scheme $\bullet$, where $k \in \{2,3\}$, $h \in \{0,1,2,3,4\}$, $\bullet \in \{\text{CHOL}, \text{TDAG}, \text{TTOP}, \text{ALG1}\}$. For a given data generating process, we calculate the following average mean squared error:

$$\widehat{\text{AMSE}}^{\bullet} = \frac{1}{10}\frac{1}{200}\sum_{k=2,3}\sum_{h=0}^{4}\sum_{l=1}^{200}(\widehat{\text{IRF}}^{\bullet}_{k,h} - \widehat{\text{IRF}}^{\text{theor.}}_{k,h})^2, \tag{24}$$

where l indicates a simulation run. We focus on the effect of the first shock to other variables since the first variable is always exogenous in all the causal structures we consider.

Table 2 shows estimated AMSE, averaged across 200 simulation experiments, for different values of T and parametrizations, along with standard errors, for the causal-chain structure. Analogous results for the common-cause (Table 3) and v-structure (Table 4) are reported in Appendix C since they are qualitatively similar.

Looking at the top panels (of all these tables), we note that the ALG1 scheme (v) outperforms the Cholesky one (scheme ii) whatever the structure and sample size when the DGP is nonlinear (second row). Recall that scheme (v) has the advantage, over scheme (ii), of allowing for nonlinearity in the estimation of impulse response functions (since it uses Gaussian process regressions via Algorithm 2), when the data are generated by a nonlinear DGP, as in this case. But it undertakes the risky task of inferring the causal order from data.[8] On the other hand, scheme (ii) runs in this case under a misspecified model (the linear one). But it has the advantage of overriding the search for a topological order from the data: it is given a correct one.

We observe a reversal in the relative accuracy of scheme (ii) and (v) when the model becomes linear (see first and third row in Table 2, subtables CHOL and ALG1). The poor performance of scheme (v) within the linear Gaussian setting does not come as a surprise since this class of ANMs is not identified. In contrast, the linear non-Gaussian model is identified, but scheme (ii) has the advantage of being endowed with the true topological order. Scheme (v) has the disadvantage of learning the causal order from data and having an estimation method which is not the most efficient for the linear case. Scheme (iv) (TTOP) fills the gap between scheme (v) and scheme (ii) in being a nonlinear approach — like the former — and hacking the order search — like the latter. As expected, scheme (iv) systematically beats scheme (v) in terms of AMSE. To assess in

[8] Peters et al. (2014) report simulation results about the accuracy of the RESIT algorithm, showing that the performance (in a nonlinear setting) in terms of distance between the estimated structure and the correct DAG is good both in outperforming other established causal discovery algorithms, but also in absolute terms. We have replicated this analysis by applying RESIT to VAR residuals and we found similar results. We also studied the accuracy of Algorithm 1 in returning a correct topological order. Both these results are available upon request. The accuracy rate is satisfactory in most cases, but we also found that small samples (i.e. $T < 500$) can be problematic for some causal structure (specifically for the common cause). Note that, however, this problem seems to have only a limited impact on the estimates of the structural IRFs (which is the main object of this study), as shown in the subsequent analysis reported in Figure 4.

Table 2: Average MSE, across 200 simulations, between IRFs estimated using different schemes (i.e. CHOL, TDAG, TTOP, ALG1) and the theoretical IRFs. MSE are calculated over the responses of $y_{t,2}$ and $y_{t,3}$ to the first shock (of one standard-deviation magnitude) and over the first five periods. The contemporaneous structure is the "causal chain". We consider different values of T (sample size) and three different model classes (Linear with Gaussian noises, non-linear with Gaussian noises and linear with non-Gaussian noises). Standard errors are reported in round brackets.

		CHOL			ALG1		
	T	250	500	1000	250	500	1000
Linear Gaussian		0.034 (1.42E-03)	0.020 (7.51E-04)	0.015 (5.20E-04)	0.257 (8.26E-03)	0.244 (7.81E-03)	0.249 (7.97E-03)
Nonlinear Gaussian		1.428 (1.05E-01)	1.023 (7.10E-02)	1.303 (8.90E-02)	0.188 (1.20E-02)	0.124 (1.00E-02)	0.085 (5.00E-03)
Linear Non-Gaussian		0.078 (3.53E-03)	0.051 (1.92E-03)	0.038 (1.82E-03)	0.293 (1.36E-02)	0.179 (8.11E-03)	0.094 (4.88E-03)
		TDAG			TTOP		
	T	250	500	1000	250	500	1000
Linear Gaussian		0.029 (1.12E-03)	0.023 (8.51E-04)	0.018 (6.44E-04)	0.036 (1.30E-03)	0.022 (8.35E-04)	0.019 (7.18E-04)
Nonlinear Gaussian		0.142 (8.00E-03)	0.086 (9.00E-03)	0.091 (8.00E-03)	0.135 (7.00E-03)	0.086 (7.00E-03)	0.081 (8.00E-03)
Linear Non-Gaussian		0.152 (6.91E-03)	0.085 (3.30E-03)	0.060 (2.70E-03)	0.230 (1.35E-02)	0.123 (6.67E-03)	0.079 (4.09E-03)

isolation the effect of the estimation method, one can compare scheme (ii) with scheme (iv) (cf. subtables CHOL and TTOP in Table 2). To assess in isolation the effect of causal learning, one can compare scheme (v) with (iv) (cf. subtables ALG1 and TTOP in Table 2). One can notice that the main source of error of scheme (v) depends on the distribution of the shocks. In the linear non-Gaussian case, the performance of scheme (v) is mostly hindered by the estimation method (AMSEs of TTOP are much closer to those of ALG1 than of CHOL). Notice that, in this setting, the estimation method improves as the sample size increases, so that the performance of scheme (iii) and (iv) gets closer to scheme (ii). In the linear Gaussian case, the performance of scheme (v) is mostly hindered by the structure learning method (AMSEs of TTOP are much closer to those of CHOL than of ALG1).

Comparisons between scheme (iv) and scheme (iii) (subtables TTOP and TDAG) do not clearly discriminate between the two. It seems that within our sample range the less parsimonious modeling strategy (scheme iv) often allows to better mimic the DGP than what can be achieved by

faithfully sticking to the actual DGP specification (scheme iii). This finding is important because it supports our decision to use Algorithm 1, i.e. a truncated version of the original RESIT by Peters et al. (2014). As regards of how AMSE changes with respect to the sample size T, we notice that with scheme (ii) AMSE does not show any downward trend over T when the DGP is nonlinear, while with scheme (v) AMSE shows a clear tendency to decrease, at least over the values of T we consider.

Figure 4 shows structural IRFs in the causal chain setting computed via Monte Carlo integration (Algorithm 2) under the theoretical scheme (i), the CHOL scheme (ii), and the ALG1 scheme (v), plotted from horizon 0 to 10. Dashed lines give the 68% confidence intervals of those averages. Specifically, the top panels show the dynamic responses of variable 2 ($y_{t,2}$) to setting structural shock $\varepsilon_{t,1}$ to one at time $t^* = 2$, while the bottom panels display their counterparts for variable 3 ($y_{t,3}$). For the sake of conciseness and readability, we do not plot IRFs obtained with TDAG and TTOP since they are very similar to each other and tend to overlap with those obtained by ALG1. For the same reasons, we display simulation results only for the smallest samples ($T = 250$, left panels) and largest ($T = 1000$, right panels). Analogous figures for the common-cause and v-structure settings can be found in Appendix C (Figures 8 and 9, respectively). From these plots, one can notice that the ALG1-IRFs are in most cases, even at horizon equal to zero, indistinguishable from the theoretical ones. Moreover, the failure of the CHOL scheme in recovering contemporaneous shocks' impacts is quite evident. Since CHOL is built on the correct (imposed a priori) topological order, this suggests that knowing this order is not enough to accurately estimate structural IRFs, but it is also crucial to allow for nonlinearities in the regression method, if present. The nonparametric regression we use (GP regression) seems to successfully capture the nonlinear contemporaneous effects.

4 Empirical application

Our empirical application studies the effects of macroeconomic shocks on nominal interest rate, inflation, output and financial conditions, using U.S. data. The study allows for nonlinearity at the contemporaneous level. We compare our findings with the results from a linear recursive SVAR model, highlighting the difference between the two results. This fact allows us to depart from a typical feature of linear SVAR models, namely the fact that the effect of a positive shock is by construction symmetric to the effect of a negative shock. Asymmetric effects of shocks have been studied, among others, by Lo and Piger (2005), Höppner et al. (2008), Kilian (2014), Hussain and Malik (2016). We should notice, however, that we are imposing a recursive structure with additive noises on the relationship among innovations, in a setting where the reduced form VAR is linear. This implies that the shock which is last in the topological order can only have a linear impact on the variables. Specifically, this means that if the monetary policy instrument is last, as it turns out to be the case in our application, its shocks have linear effects by construction.

We estimate a linear reduced-form VAR model with five variables: inflation (pi_t), output gap (out_t), the federal funds rate (r_t), industrial production (ip_t), and excess bond premium (ebp_t).

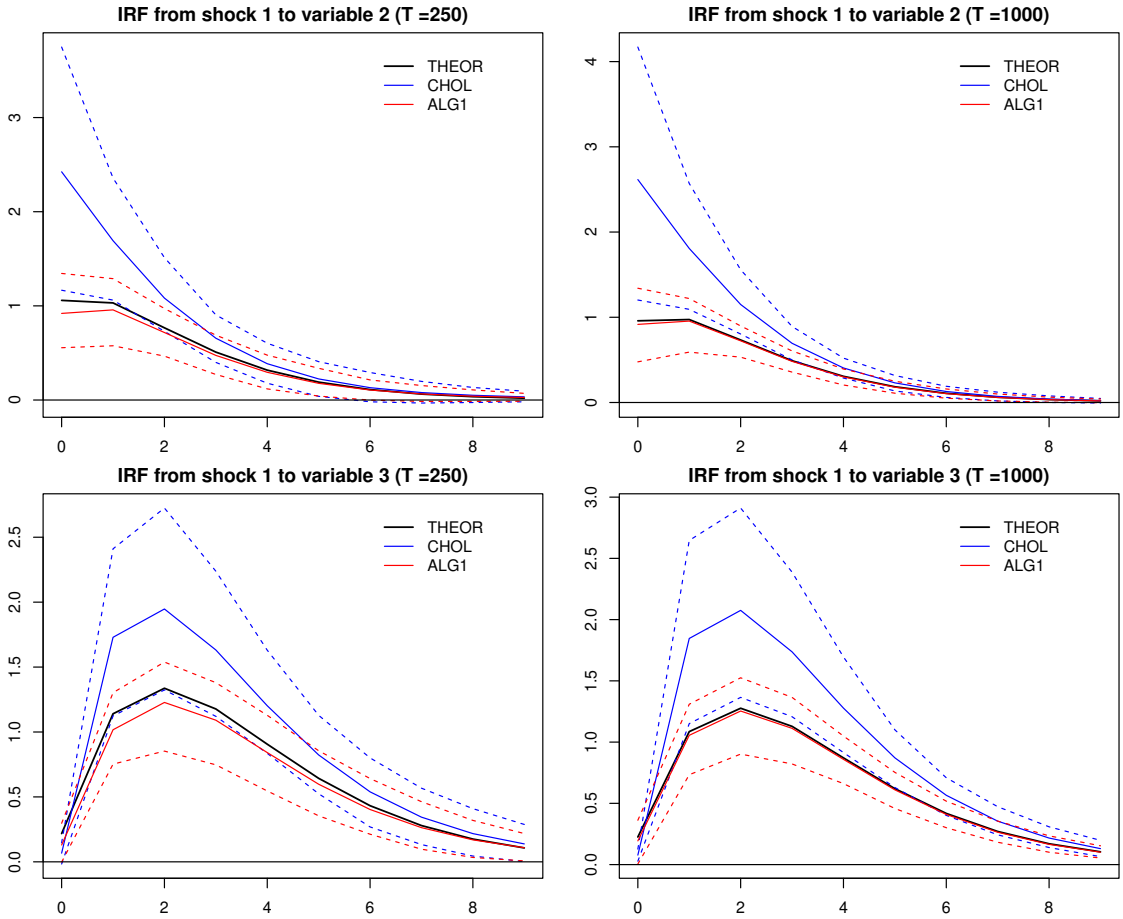

Figure 4: DGP with "causal chain". Theoretical (THEOR), linear (CHOL), and nonlinear (ALG1) IRFs for $T = 250$ (left panels) and $T = 1000$ (right panels). Each solid line exhibits the average, across 200 simulations, IRF from a unitary shock 1 ($\varepsilon_{t,1}$) to variable 2 ($y_{t+h,2}$) (top panels) or variable 3 ($y_{t+h,3}$) (bottom panels), for $h = 0, \ldots, 9$. Confidence interval at 68% are reported in dashed lines.

We take the data from the Federal Reserve Economic Database (FRED), except ebp_t, which we took from Forni et al. (2024). The data are collected at the quarterly frequency in U.S. and cover the period from 1973-Q1 to 2019-Q4 (188 observations). Inflation is computed as the change in the logarithm of the GDP deflator (FRED mnemonic GDPDEF). Output gap is computed as the deviation of the logarithm of real GDP (FRED mnemonic GDPC1) from the logarithm of potential GDP (FRED mnemonic GDPPOT), which we take from the U.S. Congressional Budget Office's estimate. The series r_t is set to the federal funds effective rate (FRED mnemonic FEDFUNDS). The series ip_t is taken from the index of industrial production (FRED mnemonic INDPRO). The five time series are plotted in Figure 5. A three-lags VAR specification is selected according to the Akaike information criterion. The "portmanteau" test for residual autocorrelation does not reject the null hypothesis of zero autocorrelation at 0.01 level of significance.

We identify a SVAR model of the type of equation (15). As mentioned in Section 2, identification can be obtained by knowledge of the causal structure among innovation terms. As noted, for

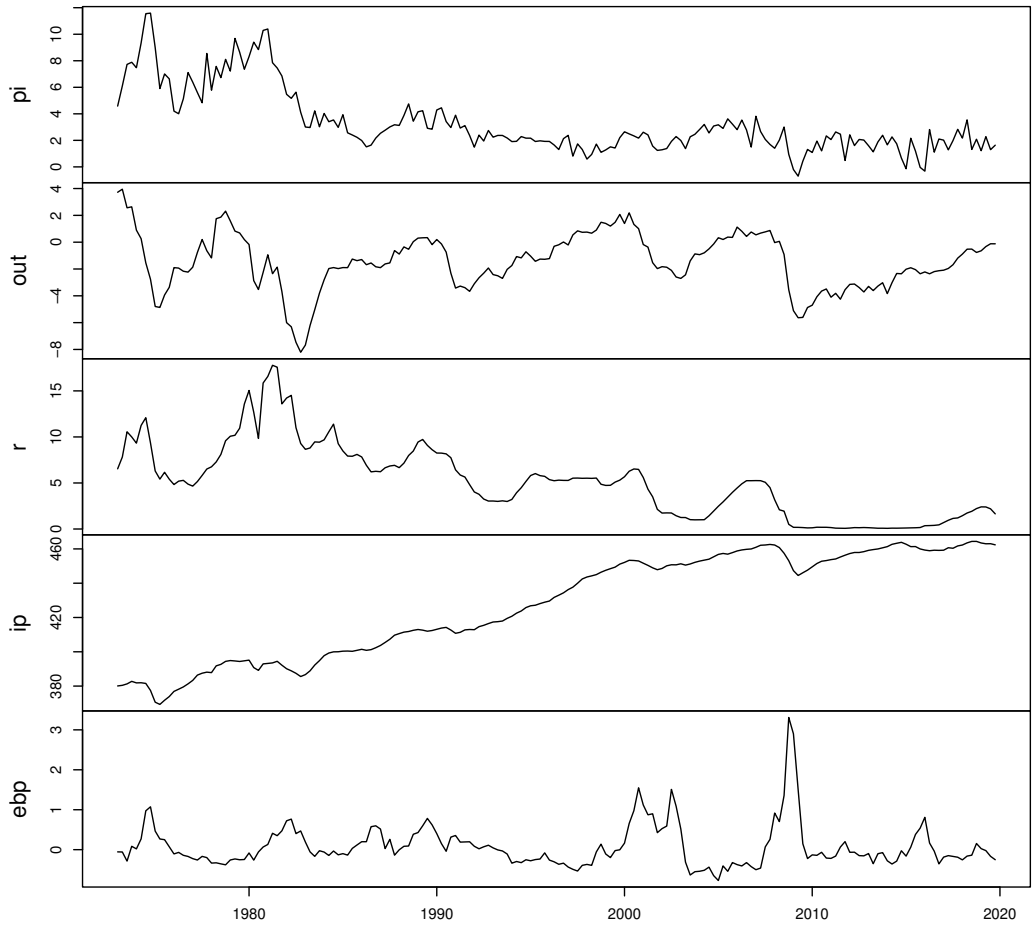

Figure 5: Employed time series. From the top, inflation (π), output gap (o) and nominal interest rate (r). Data sample from 1953-Q3 to 2019-Q4.

the sake of estimating structural impulse response functions, knowledge of the topological order is sufficient in our framework, given the recursiveness assumption. Therefore, as in the simulation analysis, we apply Algorithm 1 to the data (i.e., first step of RESIT to the reduced-form VAR residuals). Having obtained the topological order and a set of (potential) graphical parents from Algorithm 1, we feed Algorithm 2 with these two pieces of information. Thus, we can estimate the structural IRFs, setting the number of simulations for the Monte Carlo integration $N = 100$. Both in Algorithm 1 and Algorithm 2, estimates of $\varphi_k(\cdot)$ (i.e., functional dependence between innovation terms) are obtained through Gaussian process regressions.

To address uncertainty in the estimates of the structural IRFs, we perform a bootstrap analysis with 500 iterations. For each bootstrap iteration, we estimate a new VAR model and, maintaining the topological order that we got from the application of Algorithm 1 to the original data, we

compute new contemporaneous effects and structural shocks.[9] On the basis of this, we calculate structural IRFs following Algorithm 2.

The topological order that we obtain by applying Algorithm 1 to our time series is $\langle pi_t, ip_t, out_t, ebp_t, r_t \rangle$, which supports the idea that the federal funds rate is set in reaction to movements of the other variables, since the Fed is supposed to continuously monitor the macroeconomic and financial conditions of the system. Inflation, macroeconomic activity, financial conditions do not respond to changes in the federal funds rate within the period, according to our findings. This is in tune with typical recursiveness assumptions used in SVAR-based monetary analysis (see Christiano et al., 1999).

For each nonlinear IRF we analyze, we consider the response of a variable $y_{(t+h),k}$ to a shock $\varepsilon_{t,j}$ of magnitude equal to one standard deviation of the variable $y_{t,j}$.[10] We do it for $h = 0, \ldots, 20$. We compare these nonlinear IRFs with linear IRFs estimated using the following scheme: (i) we impose the Cholesky order derived from the topological order estimated by Algorithm 1 (i.e. $\langle pi_t, ip_t, out_t, ebp_t, r_t \rangle$); (ii) hence we get the structural shocks and the coefficients of the impact matrix; (iii) using the latter terms as input, we apply Algorithm 2 and get IRFs.

Figure 6 (left panels) shows the median effects of a positive interest rate shock (i.e. a contractionary monetary policy shock) on the other variables, derived from bootstrapped nonlinear and linear IRFs. Confidence intervals refer to the interquartile range (i.e. the 25th and 75th percentiles). Notice that the effects of an expansionary monetary policy shock (right panels) are by construction completely symmetric to the contractionary ones, figuring r as last in the topological order. Figure 7 shows the median effects of a positive (left panel) and negative (right panel) inflation shock. In Appendix D (see Figure 10 and Figure 11), we also show the effects of the output and the excess bond premium shocks.

As regards the r (positive) shock (Figure 6, left panel), the reaction of inflation is positive for two quarters and then becomes negative after one year. The reactions of industrial production and output gap are decisively negative. The excess bond premium responds positively, but only after 3 quarters. Notice that the nonlinear and linear scheme cannot depart from each other by construction as regards this shock.

Looking at the responses of a positive inflation shock (Figure 7, left panels), we see that industrial production has an instantaneous positive reaction. This holds both for the linear and nonlinear schemes. If we look at the responses of a negative inflation shock (Figure 7, right panels), the instantaneous effects are symmetric to the positive shock in the linear scheme by construction. But we find an asymmetry in the nonlinear scheme: the responses of ip to the negative inflation shock are also positive at the impact. In the subsequent periods, for 2-3 quarters, the effects of a positive pi shock on ip vanish and then backfire, under both the linear and nonlinear

[9] We maintain the same topological order at each bootstrap iteration because we do not want to account for uncertainty about inference of the causal structure here. To do that, one can adapt the stability selection procedure by Meinshausen and Bühlmann (2010).

[10] As starting date, we simply choose $t^* = 4$, considering that the number of lags is equal to 3. Note, however, that due to the linearity of the reduced-form model, different choices would not affect our results.

schemes. Under the latter scheme, the effects of a negative pi shock on ip remain on the positive sign. Other asymmetries are detectable also as regards the reactions of out, ebp, and r to the same shock: the instantaneous reactions of output gap and excess bond premium (in the nonlinear scheme) are not symmetric between the positive and negative pi shock. In particular, although output gap reacts instantaneously to the pi shock with the opposite sign, the absolute value of such a reaction is greater and more evident (considering the position of confidence interval above the zero line) in the case of a negative pi shock. The short-term response of the federal funds rate (r) to positive and negative inflation shocks under the nonlinear scheme is noteworthy for its asymmetry. Notably, when the system experiences a positive inflation shock, the federal funds rate reacts much more strongly, nearly doubling in absolute terms compared to its response to a negative inflation shock, as regards the point estimates. In terms of confidence intervals, the impact is significantly positive after a positive inflation shock, but not significant after a negative inflation shock. This suggest that the FED is more reactive, in terms of changing the short term interest rate, to a positive inflation shock than a negative one.

Looking at Figures 10 and 11 (Appendix D), we can detect some asymmetries in the contemporaneous responses of r to both the out and ebp shocks. In particular, looking at the nonlinear scheme, the contemporaneous median response of r is positive to out shocks, no matter the sign, and, conversely, the contemporaneous median response of r is negative to ebp shocks, no matter the sign.

5 Conclusions

There is a departure from the linear model with Gaussian disturbances that can be exploited for identification, namely nonlinearity, as long as the disturbances are additive. We have shown that this feature can be exploited to identify a structural VAR model that contains nonlinearities in the cross-sectional structure. The identification criterion we have exploited in this paper is based on the following simple idea: in case of nonlinearity it is admitted an additive noise model (i.e. a model in which disturbances are independent of covariates) from the cause to the effect, but, in the generic case (see conditions C1-C3 in section 2.3), not from the effect to the cause. Thus, by iterative regressions and subsequent independence tests, it is possible to determine the contemporaneous causal structure underlying a VAR model, under the assumption of recursiveness and mutual independence of the shocks. On this basis, we can recover structural impulse response functions, which we define as difference of conditional expectations and estimate through Monte Carlo integration. It also turns out that, for the sake of estimating impulse response functions, the inferred causal structure can contain some redundant links, as long as the topological order is correct. Simulation results have shown that the proposed search procedure and our scheme to estimate impulse response functions perform correctly under data generating processes satisfying the theoretical conditions. Moreover, if the data generating process is nonlinear, they outperform methods based on linearity assumptions, even under the correct zero restrictions. Our empirical analysis has shown that taking into account nonlinearity helps recovering the contemporaneous

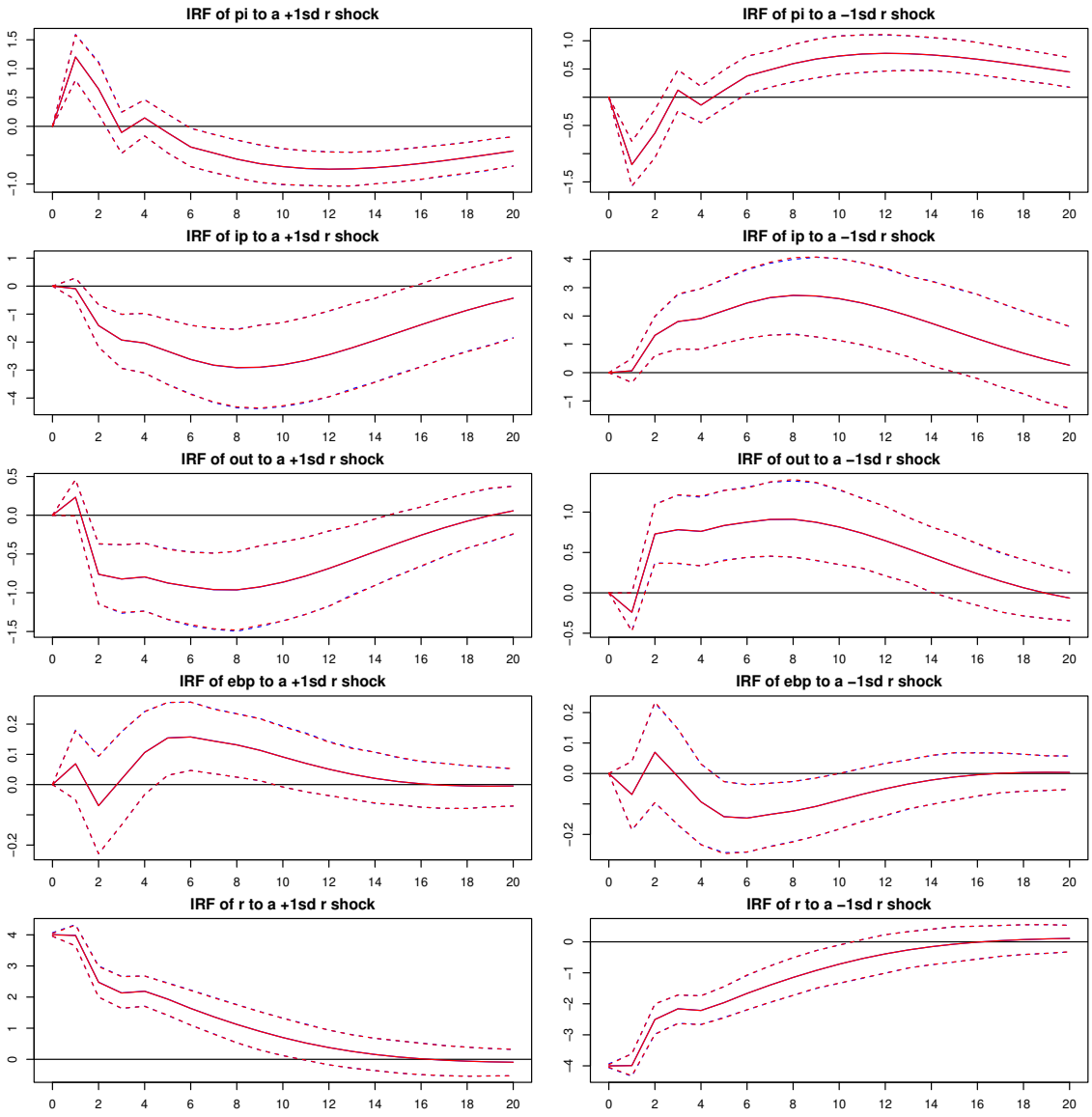

Figure 6: Structural IRFs to a positive (left panels) and negative (right panels) interest rate shock of one-standard deviation magnitude. Each panel shows nonlinear structural IRFs obtained from Algorithm 2 with GP regressions (red lines) and linear structural IRFs obtained from Algorithm 2 but fed with coefficients derived from Cholesky factorization (blue lines). Solid lines refer to the median effects while dashed lines refer to the interquartile range computed over 500 bootstrap simulations.

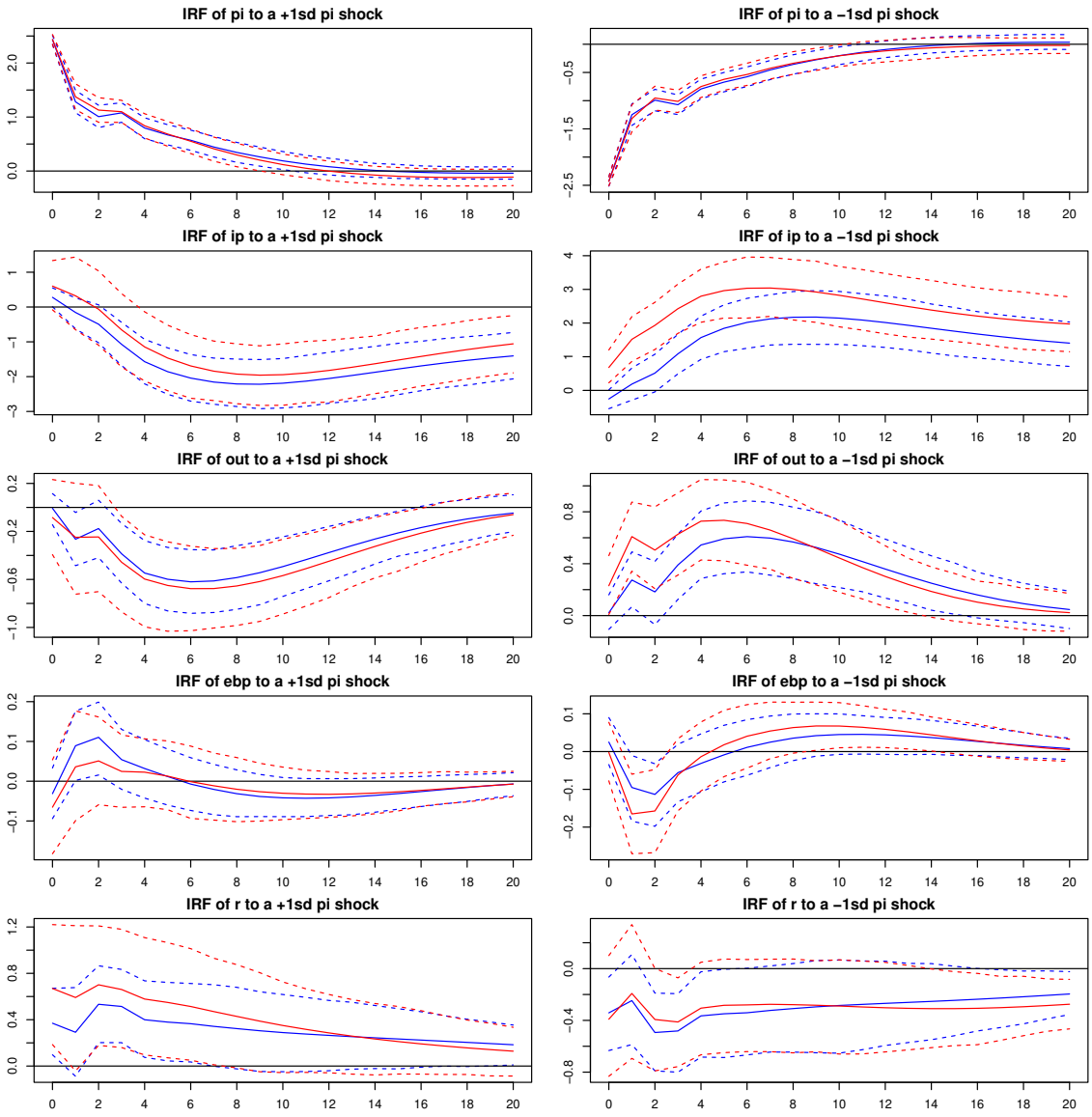

Figure 7: Structural IRFs to a positive (left panels) and negative (right panels) inflation shock of one-standard deviation magnitude. Each panel shows nonlinear structural IRFs obtained from Algorithm 2 with GP regressions (red lines) and linear structural IRFs obtained from Algorithm 2 but fed with coefficients derived from Cholesky factorization (blue lines). Solid lines refer to the median effects while dashed lines refer to the interquartile range computed over 500 bootstrap simulations.

structure of the SVAR system. It may also make a substantial difference in terms of predicting the effect of macroeconomic shocks to the economy, particularly because it allows for asymmetric effects between positive and negative structural shocks.

The idea of exploiting nonlinearity for identification has been put into practice in this paper by recovering a topological order through Algorithm 1, but one can, of course, conceive different ways to implement it. Other procedures worth exploring are those in the family of the score-based approach to causal discovery (see, e.g., Peters et al., 2014: section 4.2, and Bühlmann et al., 2014). A limitation of our procedure can be seen in the recursiveness assumption. This has been relaxed, in the bivariate case, by Mooij et al. (2011), but further research is needed for the case with more than two variables and, more in general, for its adaptation to the structural VAR framework. Another interesting path of research is the nonlinear ICA approach, which has been pioneered by Hyvärinen and Pajunen (1999) and recently developed by Monti et al. (2020) and Gunsilius and Schennach (2023). This class of blind source separation methods departs from the additive noise model framework studied in this paper and its applicability to macroeconomic data is still to be explored. Our hope is that the potentiality that nonlinearity has shown for identification will foster the adoption of this idea within the field of econometrics.

References

Bacchiocchi, E. and T. Kitagawa (2022). Locally- but not globally-identified SVARs. Working Paper, Bologna.

Blanchard, O. J. and D. Quah (1989). The dynamic effects of aggregate demand and supply disturbances. *The American Economic Review 79*(4), 655–673.

Brunnermeier, M., D. Palia, K. A. Sastry, and C. A. Sims (2021). Feedbacks: financial markets and economic activity. *American Economic Review 111*(6), 1845–79.

Bühlmann, P., J. Peters, and J. Ernest (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 2526–2556.

Choi, T. and M. J. Schervish (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis 98*(10), 1969–1987.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (1999). Monetary policy shocks: What have we learned and to what end? Volume 1 of *Handbook of Macroeconomics*, pp. 65–148. Elsevier.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing 36*(3), 287–314.

Deistler, M. and H.-G. Seifert (1978). Identifiability and consistent estimability in econometric models. *Econometrica*, 969–980.

Demiralp, S. and K. D. Hoover (2003). Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics 65*, 745–767.

Eriksson, J. and V. Koivunen (2004). Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters 11*(7), 601–604.

Fan, J. and I. Gijbels (2018). *Local polynomial modelling and its applications*. Routledge.

Faust, J. (1998). The robustness of identified VAR conclusions about money. In *Carnegie-Rochester conference series on public policy*, Volume 49, pp. 207–244. Elsevier.

Fiorentini, G. and E. Sentana (2023). Discrete mixtures of normals pseudo maximum likelihood estimators of structural vector autoregressions. *Journal of Econometrics 235*(2), 643–665.

Forni, M., L. Gambetti, N. Maffei-Faccioli, and L. Sala (2024). Nonlinear transmission of financial shocks: Some new evidence. *Journal of Money, Credit and Banking 56*(1), 5–33.

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics 8*(2), 261–263.

Gouriéroux, C., A. Monfort, and J.-P. Renne (2017). Statistical inference for independent component analysis: Application to structural VAR models. *Journal of Econometrics 196*(1), 111–126.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola (2007). A kernel statistical test of independence. *Advances in neural information processing systems 20*.

Guay, A. (2021). Identification of structural vector autoregressions through higher unconditional moments. *Journal of Econometrics 225*(1), 27–46.

Guerini, M. and A. Moneta (2017). A method for agent-based models validation. *Journal of Economic Dynamics and Control 82*, 125–141.

Gunsilius, F. and S. Schennach (2023). Independent nonlinear component analysis. *Journal of the American Statistical Association 118*(542), 1305–1318.

Györfi, L., M. Kohler, A. Krzyzak, H. Walk, et al. (2002). *A distribution-free theory of nonparametric regression*, Volume 1. Springer.

Härdle, W., H. Lütkepohl, and R. Chen (1997). A review of nonparametric time series analysis. *International statistical review 65*(1), 49–72.

Hastie, T. J. and R. J. Tibshirani (2017). *Generalized additive models*. Routledge.

Herwartz, H. (2018). Hodges–Lehmann detection of structural shocks–an analysis of macroeconomic dynamics in the Euro area. *Oxford Bulletin of Economics and Statistics*.

Herwartz, H., A. Lange, and S. Maxand (2022). Data-driven identification in SVARs—when and how can statistical characteristics be used to unravel causal relationships? *Economic Inquiry 60*(2), 668–693.

Höppner, F., C. Melzer, and T. Neumann (2008). Changing effects of monetary policy in the us–evidence from a time-varying coefficient var. *Applied Economics 40*(18), 2353–2360.

Hoyer, P., D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems 21*, 689–696.

Hussain, S. M. and S. Malik (2016). Asymmetric effects of exogenous tax changes. *Journal of Economic Dynamics and Control 69*, 268–300.

Hyvärinen, A. (2013). Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1984), 20110534.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent component analysis*. John Wiley & Sons.

Hyvärinen, A. and P. Pajunen (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks 12*(3), 429–439.

Kilian, L. (2014). Oil price shocks: Causes and consequences. *Annu. Rev. Resour. Econ. 6*(1), 133–154.

Kilian, L. and H. Lütkepohl (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

Koop, G., M. H. Pesaran, and S. M. Potter (1996). Impulse response analysis in nonlinear multivariate models. *Journal of econometrics 74*(1), 119–147.

Koopmans, T. C. and O. Reiersøl (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics 21*(2), 165–181.

Lanne, M. and J. Luoto (2021). GMM estimation of non-gaussian structural vector autoregression. *Journal of Business & Economic Statistics 39*(1), 69–81.

Lanne, M. and H. Lütkepohl (2008). Identifying monetary policy shocks via changes in volatility. *Journal of Money, Credit and Banking 40*(6), 1131–1149.

Lanne, M. and H. Lütkepohl (2010). Structural vector autoregressions with nonnormal residuals. *Journal of Business & Economic Statistics 28*(1), 159–168.

Lanne, M., M. Meitz, and P. Saikkonen (2017). Identification and estimation of non-Gaussian structural vector autoregressions. *Journal of Econometrics 196*(2), 288–304.

Lewis, D. J. (2021). Identifying shocks via time-varying volatility. *The Review of Economic Studies*.

Lo, M. C. and J. Piger (2005). Is the response of output to monetary policy asymmetric? Evidence from a regime-switching coefficients model. *Journal of Money, credit and Banking*, 865–886.

Matzkin, R. L. (2007). Nonparametric identification. *Handbook of econometrics 6*, 5307–5368.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Mesters, G. and P. Zwiernik (2022). Non-independent components analysis. *arXiv preprint arXiv:2206.13668*.

Moneta, A. (2008). Graphical causal models and VARs: An empirical assessment of the real business cycles hypothesis. *Empirical Economics 35*(2), 275–300.

Moneta, A., D. Entner, P. O. Hoyer, and A. Coad (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics 75*(5), 705–730.

Monti, R. P., K. Zhang, and A. Hyvärinen (2020). Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pp. 186–195. PMLR.

Montiel Olea, J. L., M. Plagborg-Møller, and E. Qian (2022). SVAR identification from higher moments: Has the simultaneous causality problem been solved? *AEA Papers and Proceedings 112*, 481–85.

Montiel Olea, J. L., J. H. Stock, M. W. Watson, et al. (2021). Inference in structural vector autoregressions identified with an external instrument. *Journal of Econometrics 225*(1), 74–87.

Mooij, J. M., D. Janzing, T. Heskes, and B. Schölkopf (2011). On causal discovery with cyclic additive noise models. *Advances in neural information processing systems 24*.

Pearl, J. (2009). *Causality. Models, Reasoning, and Inference*. Cambridge University Press.

Peters, J., D. Janzing, and B. Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Peters, J., J. Mooij, D. Janzing, and B. Schölkopf (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research 15*(1), 2009–2053.

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. *Handbook of Macroeconomics 2*, 71–162.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 375–389.

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics 85*(4), 777–792.

Roehrig, C. S. (1988). Conditions for identification in nonparametric and parametric models. *Econometrica*, 433–447.

Romer, C. D. and D. H. Romer (1989). Does monetary policy matter? a new test in the spirit of friedman and schwartz. *NBER macroeconomics annual 4*, 121–170.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 577–591.

Rubio-Ramirez, J. F., D. F. Waggoner, and T. Zha (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies 77*(2), 665–696.

Sentana, E. and G. Fiorentini (2001). Identification, estimation and testing of conditionally heteroskedastic factor models. *Journal of Econometrics 102*(2), 143–164.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, and A. Kerminen (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research 7*, 2003–2030.

Shimizu, S., T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research 12*, 1225–1248.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 1–48.

Sims, C. A. (2020). SVAR identification through heteroskedasticity with misspecified regimes. *Princeton University*.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (2000). *Causation, prediction, and search*. MIT press.

Stock, J. H. and M. W. Watson (2001). Vector autoregressions. *Journal of Economic perspectives 15*(4), 101–115.

Swanson, N. R. and C. W. Granger (1997). Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association 92*(437), 357–367.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning*, Volume 2. MIT press Cambridge, MA.

Wold, H. O. (1960). A generalization of causal chain models. *Econometrica*, 443–463.

Zhang, K. and A. Hyvärinen (2009). On the identifiability of the post-nonlinear causal model. *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 647–655.

Appendix

A Complete RESIT algorithm

Algorithm 1[full] VAR + RESIT (Phases 1-2 are from Peters et al., 2014)

Input: A K-dimensional time series vector $(y_{t,1}, \ldots, y_{t,K})'$
PHASE 0: Estimate the reduced-form model.
Estimate a reduced-form time series model of the class GVAR-ANIM, see eq. (1) and extract residuals $\widehat{u}_t = (\widehat{u}_{t,1}, \ldots, \widehat{u}_{t,K})'$. To simplify the notation, let us call $\widehat{u}_{t,k} \equiv v_k$, for $k = 1, \ldots, K$.
$S := 1, ..., K$, $\pi := [\quad]$

PHASE 1: Determine topological order.
repeat
    for $k \in S$ **do**
        Regress v_k on $\{v_i\}_{i \in S \setminus \{k\}}$ and obtain residuals e_k
        Measure dependence between e_k and $\{v_i\}_{i \in S \setminus \{k\}}$
    end for
    Let k^* be the k with the weakest dependence
    $S := S \setminus k^*$
    $pa(k^*) := S$
    $\pi := [k^*, \pi]$
until $\#S = 0$
Output: π

PHASE 2 (*pruning*): Remove superfluous edges.
for $k \in \{2, \ldots, K\}$ **do**
    for $p \in pa(\pi(k))$ **do**
        Regress $v_{\pi(k)}$ on $\{v_i\}_{i \in pa(\pi(k)) \setminus \{p\}}$.
        if residuals are independent of $\{v_i\}_{i \in \{\pi(1), \ldots, \pi(k-1)\}}$ **then**
            $pa(\pi(k)) := pa(\pi(k)) \setminus \{p\}$
        end if
    end for
end for
Output: $(pa(1), \ldots, pa(K))$

Note: with $pa(k)$ we refer to the set of indices associated to the variables in $Pa(v_k)$, for any k.

B Discussion on consistency

The method proposed provides the topological order and causal structure among innovation terms and allows us to estimate structural impulse response functions. As one gets more data, one would hope that the recovered order/structure is correct and that the estimated IRFs would converge to the true IRFs, under the assumptions that the underlying data generating process is GVAR-ANIM and that the relationships between innovation terms (equation 3) can be described by a multivariate ANM under assumptions (C1)-(C3). While it is clear that an unidentified model cannot be consistently estimated, it is not necessarily true that identifiability implies consistency (see Deistler and Seifert, 1978; Gabrielsen, 1978). Studying consistency and other statistical properties (e.g. asymptotic normality, efficiency) of the entire procedure, considering the different options in terms of model specification and estimation, is a task we leave for future research. In the following, however, we refer to some key results for such task.

1. Assume a joint distribution $P(u)$ generated by a GVAR-ANIM, associated to a DAG $\mathcal{G}$ representing the structural relations between innovation terms $u_t = (u_{t,1}, \ldots, u_{t,K})'$. Assume that: (i) conditions C1-C3 are satisfied; (ii) the u_t are consistently estimated from the reduced-form VAR model (Alg. 1, Phase 0); (iii) $\varphi_k(\cdot)$ is estimated with a consistent regression method, for each k; (iv) the dependence measures (Alg. 1, Phase 1) is correct and independence tests (Alg. 1, Phase 2, see Appendix A) are consistent. Then the output of the (full) Algorithm 1 is guaranteed to be correct. This statement generalizes Theorem 34 in Peters et al. (2014) to the case of VAR-estimated input and presupposes the identifiability result of Theorem 3 (section 2.3).

2. The regression method to estimate $\varphi_k(\cdot)$ is user-specified in our procedure, with the caveat that linear regression is not allowed, except in the case in which the structural shocks are non-normal. In our simulation analysis and empirical application, we use Gaussian process regressions, whose consistency has been studied by Williams and Rasmussen (2006, Chapter 7) in the general case and by Choi and Schervish (2007) in the univariate case. Following the latter authors, let us consider $Y_i = \eta(X_i) + \epsilon_i$ $(i = 1, \ldots, n)$, where we are interested in estimating $\eta(x) = E[Y|X = x]$. Assume: (i) $\epsilon_i \sim N(0, \sigma^2)$ or $DE(0, \sigma)$ (DE stands for double exponential, i.e. Laplace, distribution); (ii) $\eta(\cdot)$ is a random process with a Gaussian process prior distribution: $\eta(\cdot) \sim GP(\mu(\cdot), R(\cdot, \cdot))$; (iii) smoothness conditions on $\mu(\cdot), R(\cdot, \cdot))$, and $\eta(\cdot)$ (in particular, continuously differentiable mean function $\mu(x)$ and covariance function $R(\cdot, \cdot)$ having continuous fourth partial derivatives). Let the posterior distribution $p_{n,N}(\eta|X, Y)$ be, for each neighborhood N of the true regression function η_0 and each sample size n: $p_{n,N} = \Pr(\{\eta \in N\}|X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. Under the assumptions above, Choi and Schervish (2007) show that the posterior distribution of η is almost surely consistent, i.e., for every N, $\lim_{n\to\infty} p_{n,N} = 1$ a.s..

As an alternative nonparametric regression method to estimate $\varphi_k(\cdot)$, let us consider the Nadaraya-Watson estimator. Considering again the regression function $\eta(x)$, this estimator computes $\hat{\eta}(x) = \sum_{i=1}^{n} w_i Y_i$ where $w_i = \frac{K_h(X_i - x)}{\sum_{j=1}^{n} K_h(X_j - x)}$, being K_h a kernel function with bandwidth h. It has been established (see, e.g., Györfi et al., 2002) that if $nh \to \infty$ and $n \to \infty$ then the estimator is con-

sistent (the result has also been extended to the multivariate case).

3. Conditions for a consistent estimation of a fully non-parametric (but time invariant) reduced-form GVAR-ANIM (see equation 1) are discussed in Härdle et al. (1997). Nonparametric kernel estimation turns out to be consistent under the conditions discussed above for the Nadaraya-Watson estimator with the addition that the underlying process must be α- or ϕ-mixing. In our simulation and empirical application, we actually estimate a linear reduced-form VAR. Here the least squares estimator is consistent and asymptotic normal (even in the presence of unit roots with a number of lags strictly greater than one) under general conditions (see Kilian and Lütkepohl, 2017). Notice, however, that we cannot assume normality of the reduced-form residuals here, since this would imply a linear form in the conditional expectation function. Conditions for consistency in the case of a linear non-Gaussian SVAR are studied by Gouriéroux et al. (2017).

4. The dependence measure and the independence test used in the Algorithm 1 (full version, see Appendix A) are also user-specified in our procedure, under the condition that they account for order statistics. In our applications, we use the nonparametric measure of dependence and the kernel independence test proposed by Gretton et al. (2007), who prove consistency of the test by deriving distribution of the test both under the null hypothesis of independence and under the alternative. The distribution of the test statistics, as the sample size approaches infinity, is parameterised in terms of kernels of the data.

C Further simulation results: common-cause and v- structure

Table 3: Average MSE, across 200 simulations, between IRFs estimated using different schemes (i.e. CHOL, TDAG, TTOP, ALG1) and the theoretical IRFs. MSE are calculated over the responses of $y_{t,2}$ and $y_{t,3}$ to the first shock (of one standard-deviation magnitude) and over the first five periods. The contemporaneous structure is the "common cause". We consider different values of T (sample size) and three different model classes (Linear with Gaussian noises, non-linear with Gaussian noises and linear with non-Gaussian noises). Standard errors are reported in round brackets.

		CHOL			ALG1		
	T	250	500	1000	250	500	1000
Linear Gaussian		0.031 (1.29E-03)	0.019 (7.05E-04)	0.014 (4.74E-04)	0.200 (7.27E-03)	0.209 (7.40E-03)	0.189 (7.22E-03)
Nonlinear Gaussian		1.403 (1.04E-01)	1.01 (7.00E-02)	1.28 (8.80E-02)	0.185 (1.20E-02)	0.12 (9.00E-03)	0.077 (4.00E-03)
Linear Non-Gaussian		0.068 (3.09E-03)	0.045 (1.74E-03)	0.035 (1.58E-03)	0.268 (1.22E-02)	0.162 (7.12E-03)	0.089 (4.51E-03)
		TDAG			TTOP		
	T	250	500	1000	250	500	1000
Linear Gaussian		0.028 (1.01E-03)	0.023 (8.27E-04)	0.018 (6.16E-04)	0.031 (1.01E-03)	0.019 (6.85E-04)	0.017 (6.20E-04)
Nonlinear Gaussian		0.155 (8.00E-03)	0.093 (9.00E-03)	0.094 (8.00E-03)	0.14 (7.00E-03)	0.089 (6.00E-03)	0.081 (8.00E-03)
Linear Non-Gaussian		0.179 (9.27E-03)	0.102 (4.80E-03)	0.069 (3.30E-03)	0.182 (1.04E-02)	0.102 (5.37E-03)	0.070 (3.52E-03)

Table 4: Average MSE, across 200 simulations, between IRFs estimated using different schemes (i.e. CHOL, TDAG, TTOP, ALG1) and the theoretical IRFs. MSE are calculated over the responses of $y_{t,2}$ and $y_{t,3}$ to the first shock (of one standard-deviation magnitude) and over the first five periods. The contemporaneous structure is the "v-structure". We consider different values of T (sample size) and three different model classes (Linear with Gaussian noises, non-linear with Gaussian noises and linear with non-Gaussian noises). Standard errors are reported in round brackets.

		CHOL			ALG1		
	T	250	500	1000	250	500	1000
Linear Gaussian		0.019 (8.23E-04)	0.010 (4.23E-04)	0.008 (2.97E-04)	0.084 (4.56E-03)	0.089 (5.01E-03)	0.094 (5.11E-03)
Nonlinear Gaussian		0.936 (9.90E-02)	0.678 (6.80E-02)	0.861 (8.50E-02)	0.118 (1.10E-02)	0.064 (5.00E-03)	0.048 (4.00E-03)
Linear Non-Gaussian		0.057 (2.79E-03)	0.034 (1.42E-03)	0.023 (1.26E-03)	0.255 (1.40E-02)	0.135 (7.07E-03)	0.057 (3.44E-03)
		TDAG			TTOP		
	T	250	500	1000	250	500	1000
Linear Gaussian		0.016 (7.14E-04)	0.012 (5.00E-04)	0.009 (3.65E-04)	0.023 (9.47E-04)	0.014 (5.73E-04)	0.011 (4.37E-04)
Nonlinear Gaussian		0.088 (7.00E-03)	0.049 (6.00E-03)	0.053 (7.00E-03)	0.089 (7.00E-03)	0.053 (6.00E-03)	0.046 (5.00E-03)
Linear Non-Gaussian		0.121 (8.11E-03)	0.064 (3.94E-03)	0.049 (3.19E-03)	0.207 (1.26E-02)	0.102 (5.60E-03)	0.066 (3.95E-03)

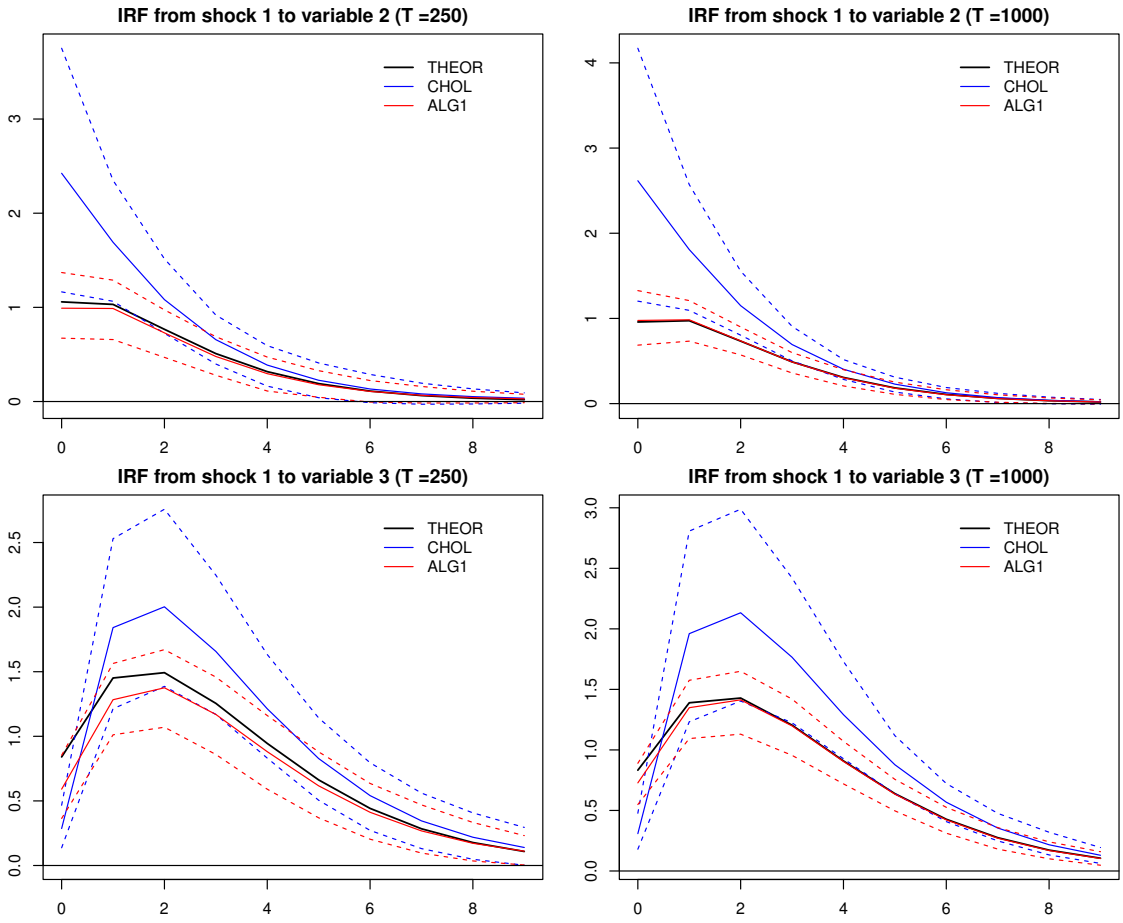

Figure 8: DGP with "common cause". Theoretical (THEOR), linear (CHOL), and nonlinear (ALG1) IRFs for $T = 250$ (left panels) and $T = 1000$ (right panels). Each solid line exhibits the average, across 200 simulations, IRF from a unitary shock 1 ($\varepsilon_{t,1}$) to variable 2 ($y_{t+h,2}$) (top panels) or variable 3 ($y_{t+h,3}$) (bottom panels), for $h = 0, \ldots, 9$. Confidence interval at 68% are reported in dashed lines.

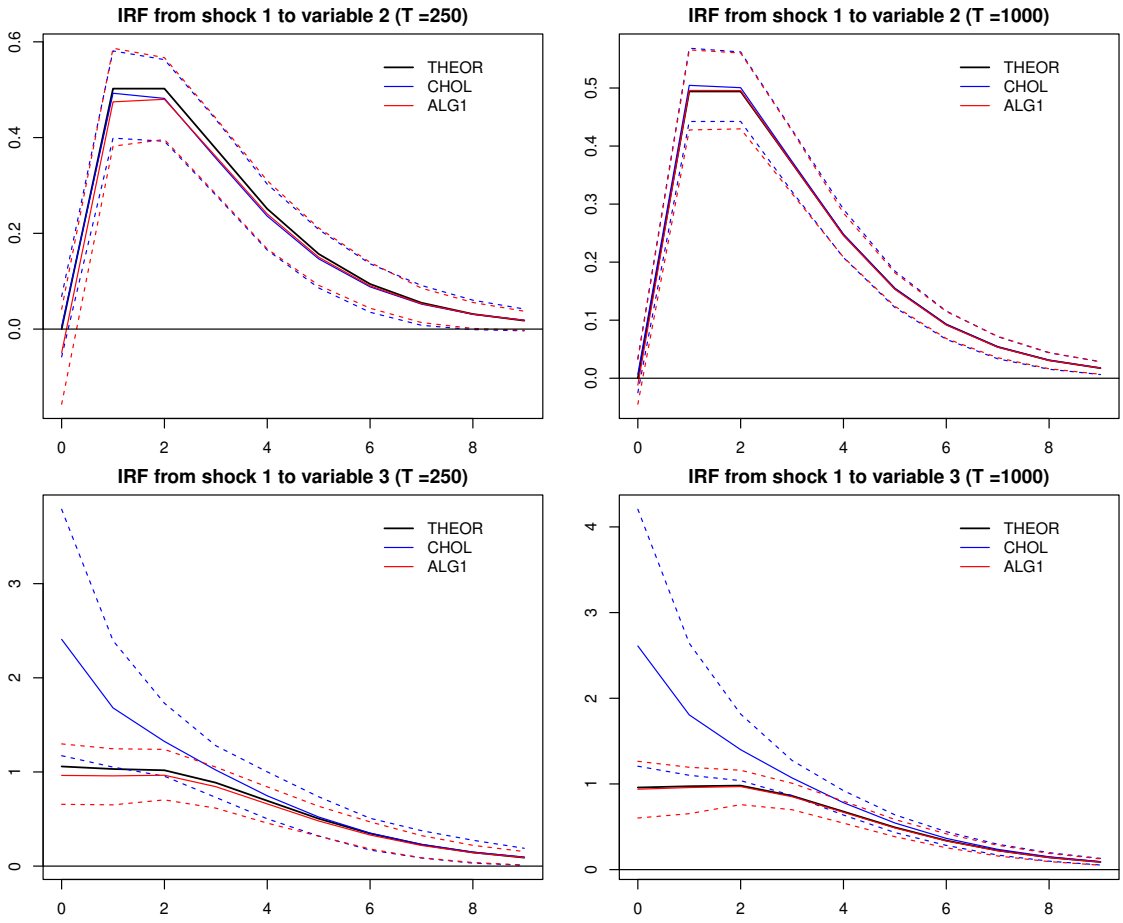

Figure 9: DGP with "v-structure". Theoretical (THEOR), linear (CHOL), and nonlinear (ALG1) IRFs for $T = 250$ (left panels) and $T = 1000$ (right panels). Each solid line exhibits the average, across 200 simulations, IRF from a unitary shock 1 ($\varepsilon_{t,1}$) to variable 2 ($y_{t+h,2}$) (top panels) or variable 3 ($y_{t+h,3}$) (bottom panels), for $h = 0, \ldots, 9$. Confidence interval at 68% are reported in dashed lines.

D Empirical results: further IRFs

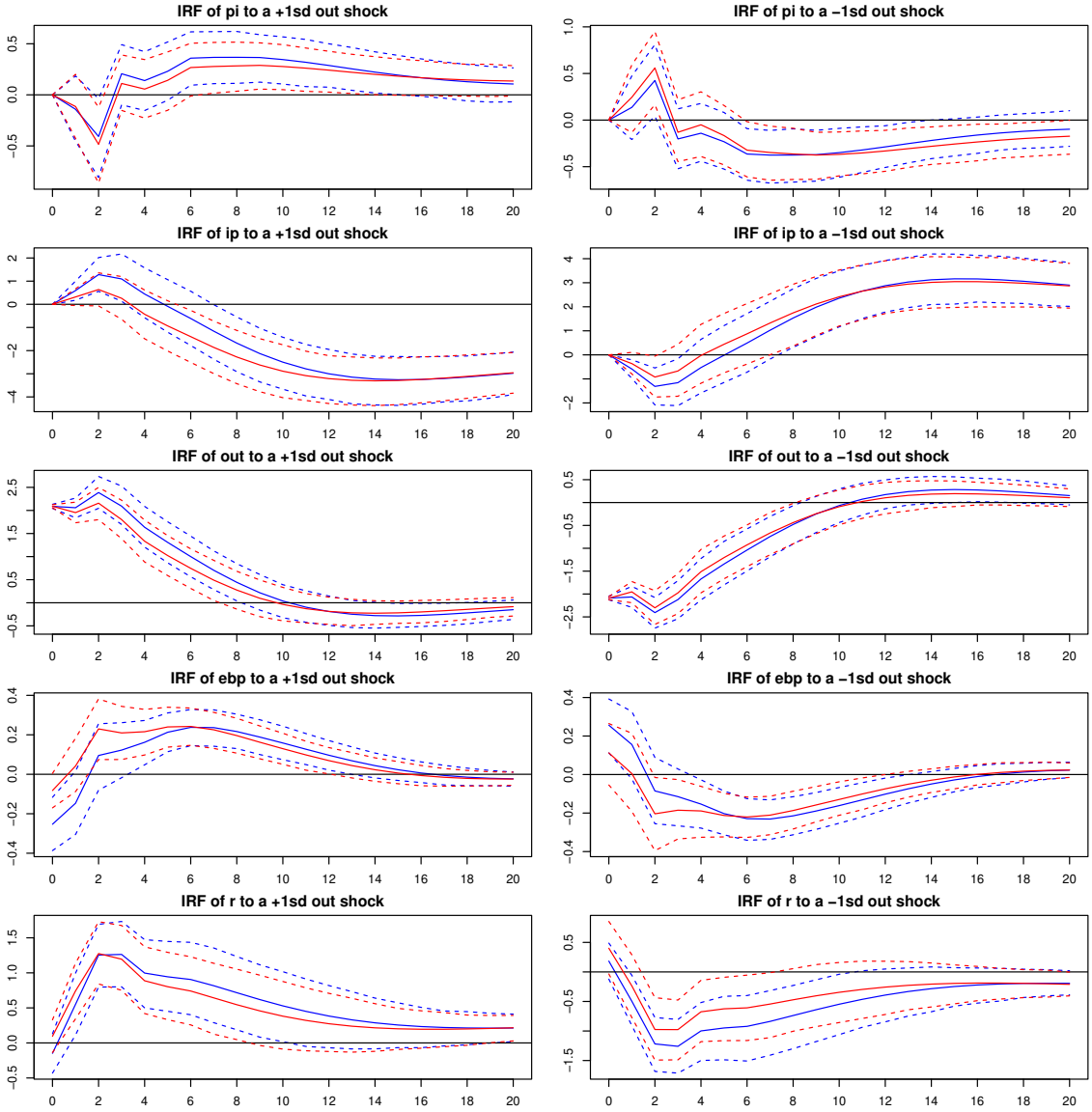

Figure 10: Structural IRFs to a positive (left panels) and negative (right panels) output gap shock of one-standard deviation magnitude. Each panel shows nonlinear structural IRFs obtained from Algorithm 2 with GP regressions (red lines) and linear structural IRFs obtained from Algorithm 2 but fed with coefficients derived from Cholesky factorization (blue lines). Solid lines refer to the median effects while dashed lines refer to the interquartile range computed over 500 bootstrap simulations.

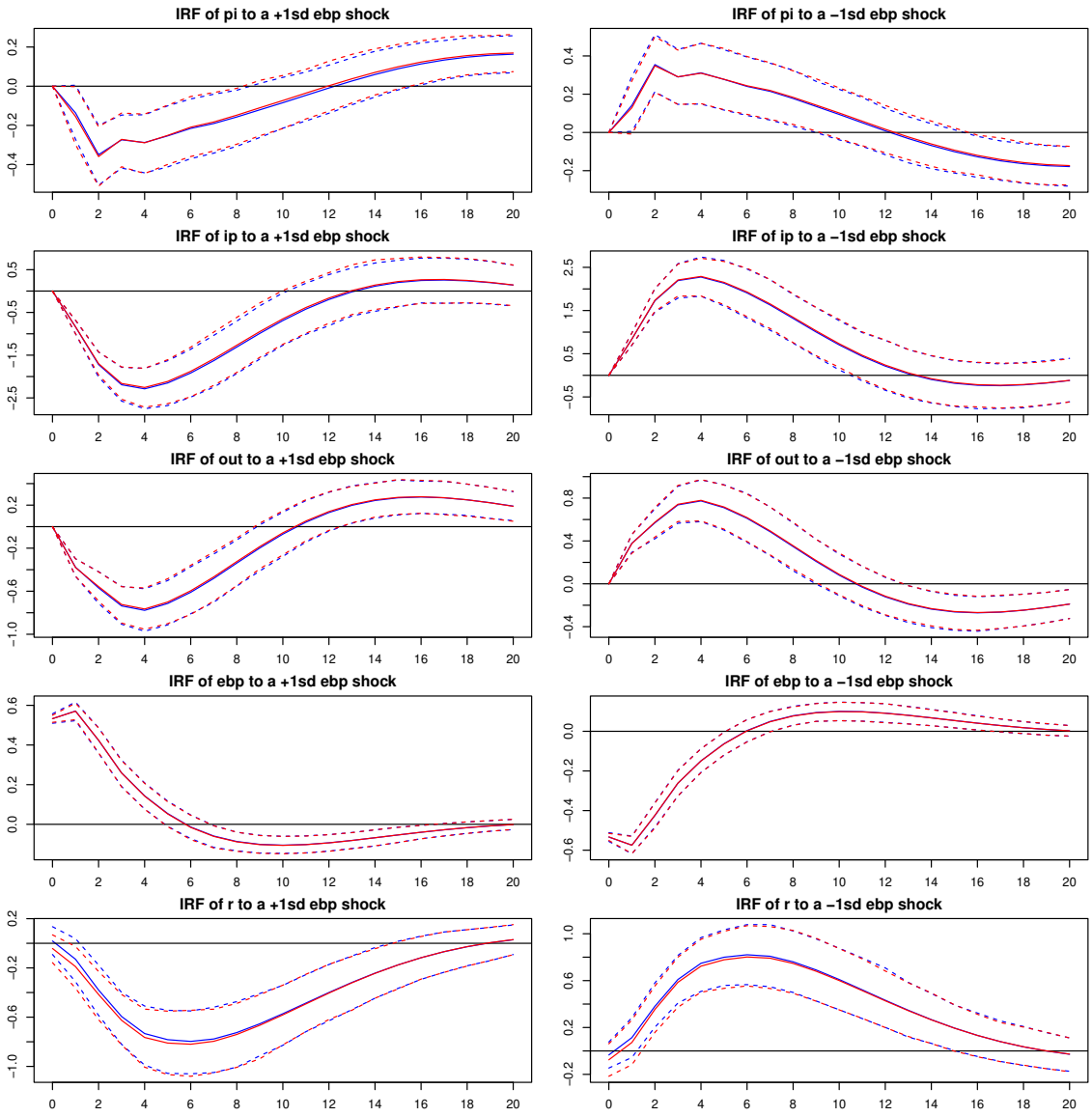

Figure 11: Structural IRFs to a positive (left panels) and negative (right panels) excess bond premium shock of one-standard deviation magnitude. Each panel shows nonlinear structural IRFs obtained from Algorithm 2 with GP regressions (red lines) and linear structural IRFs obtained from Algorithm 2 but fed with coefficients derived from Cholesky factorization (blue lines). Solid lines refer to the median effects while dashed lines refer to the interquartile range computed over 500 bootstrap simulations.