Gaussian Copulas for Imposing Structure on VAR

Aramayis Dallakyan and David A. Bessler

Department of Agriculture Economics, Texas A&M University

Selected Paper prepared for presentation at the 2018 Agricultural & Applied Economics Association Annual Meeting, Washington, D.C., August 5-August 7

Gaussian Copulas for Imposing Structure on VAR

Aramayis Dallakyan David A. Bessler

Department of Agriculture Economics, Texas A&M University

May 23, 2018

Abstract

The vector autoregression (VAR) model profoundly uses the lagged causal relationships among variables. It is well known that VAR models say little about contemporaneous time correlation of these variables. However, ignoring causal orderings among a VAR endogeneous variables in contemporaneous time may produce not representative impulse response simulations and Forecast Error Variance (FEV) decomposition. The recent advances in Machine Learning and Statistical Learning literature allow researchers to use Directed Acyclic Graphs(DAGs) to discover causal relationship from the data and help to impose structure on VAR. In this paper, we propose extended version of using DAGs to impose structure on VAR when the data does not follow normal distribution. We show the performance of our method using high dimensional simulation studyand as well real Macreconomic data.

Keywords: Graphical Models, Lasso VAR, SVAR

1 Introduction

The vector autoregression (VAR) model profoundly uses the lagged causal relationships among variables. It is well known that VAR models say little about contemporaneous time correlation of these variables. However, ignoring causal orderings among a VAR endogeneous variables in contemporaneous time may produce not representative impulse response simulations and Forecast Error Variance (FEV) decomposition (Bessler, 1984; Sims, 1980).

Econometric literature for VAR's has traditionally accounted for contemporaneous correlation in several ways. First is the Cholesky factorization, where contemporaneous correlations are established by imposing theoretically based and recursive causal ordering on the VARs variance/covariance matrix. The problem with a first approach is that impulse response and Forecast Error Variance decomposition results vary with the ordering chosen by Cholesky-factorization. The second approach is Bernanke (1986) structural VAR (SVAR) methods, where prior notions of evidentially based and/or theoretically grounded contemporaneously causal orderings may be imposed on a VARs endogenous variables. The problem here is that the true contemporaneous orderings that the researcher claims to know may be in fact unknown. The solution to this problem, that is making inference about causal ordering from the data, was first given in a literature, known as the graph-theoretic approach to causal inference by Pearl (2009); Spirtes et al. (2000). The early users that use this methods (mainly PC [1] algorithm) to solve the problem of determining the causal order of the structural VAR were Swanson and Granger (1997); Bessler and Akleman (1998); Bessler and Lee (2002); Demiralp and Hoover (2003). . By following this procedure researchers avoid choosing arbitrarily among competing but otherwise theoretically consistent sets of contemporaneous orderings inherent in Choleski-factorization or Bernanke structural VARs.

The previous use of graph theoretical methods for SVAR were limited only for Gaussian distributions and low dimensional series. (Bessler and Akleman, 1998; Haigh and Bessler, 2004; Demiralp and Hoover, 2003; Demiralp et al., 2014). In this paper we extend the existing method to a high-dimensional consistent method for a broader class of Gaussian copula or nonparanormal distribution (Liu et al., 2009, 2012; Harris and Drton, 2013) using rank-based measures of correlation such as Kendalls τ and Spearmans ρ. We call this procedure NPNDAG (Nonparanormal Directed Acyclic

[1] Named after authors Spirtes et al. (2000)

Graph). Following the same logic, for making the reference easier we called the procedure define by Swanson and Granger (1997); Bessler and Akleman (1998) as PCDAG.

Since we will implement NPNDAG in high-dimensional environment than we need VAR model that designed to solve high dimensional problems. Unfortunately, it is well known that traditional VAR is not designed for solving high-dimensional problems, since the growth of a number of parameters in a $p - lag$ VAR with the number of component series (K) is quadratic.(K^2p). For an even moderate dimensional model this can lead to noisy AR parameter estimates. To overcome this drawback, many sparse VAR approaches have been proposed, including factor models (Forni et al., 2000; Stock and Watson, 2002), where the authors assume the interdependence within high dimensional data can be explained by a few common factors. Banbura et al. (2010) and Koop (2011) used a Bayesian approach for large VAR. They found that applying Bayesian shrinkage using priors as proposed by Doan et al. (1984); Litterman (1986) is sufficient to deal with large models, provided that the tightness of the priors should be increased as more variables are added. Many recent popular approaches are based on advances in variable selection (Tibshirani, 1996) and its variants (Zhao and Yu, 2006; Yuan and Lin, 2006), use sparsity penalty for the AR coefficients. For example, see Shojaie and Michailidis (2010); Song and Bickel (2011); Nicholson et al. (2016). The advantage of the Lasso-VAR approach is that the model selection and parameter estimation are conducted simultaneously. However, there are also several disadvantages with this approach. One of the serious disadvantage is , because of its nature, the VAR model is represented as linear regression model and the current values of the time series are referred as the response variable and lagged values as the explanatory variables. However, defining the problem in this way forces the model to ignore the temporal dependence in the time series. Song and Bickel (2011); Nicholson et al. (2016) give several solutions and theoretical discussion on the consequences of not accounting for temporal dependence into the model.To mitigate this issue, Nicholson et al. (2016) suggested the hierarchical vector autoregression (HVAR) framework, which offers three different structures to allow for varying levels of flexibility. Other approaches to introduce sparsity in a VAR (sVAR) has been proposed by Davis et al. (2016), where they developed a two-stage approach of fitting sVAR models. The first stage selects nonzero AR coefficients by analyzing marginal series that are conditionally correlated. The conditional correlation between marginal time series is computed

by partial spectral coherence (PSC). PSC is a convenient measure in frequency-domain time series analysis that can be used to estimate conditional dependence between components of a multivariate time series. Since the VAR model in stage 1 may contain spurious nonzero AR coefficients, stage 2 uses t-ratios of the AR coefficient estimates to refine the model.

In this paper,first we conduct simulations to estimate the performance of NPNDAG method based on ability to recover true causal ordering among VAR endogenous variables under different distributional assumptions and sparsity levels. Then proposed method was applied to Stock and Watson (2005) data to illustrate empirical usefulness using residuals from high dimensional VAR methods such as HVAR.

The remainder of this paper is organized as follows. In Section 2 we review some material on SVAR and its identification as well we give overview of HVAR model. In Section 3 we review graphical models,PC algorithm and describe nonparanormal distribution and extended PC algorithm for Gaussian Copula based graphical models. A simulation study in Section 4 and real world data application in Section 5 emphasizes the advantage of NPNDAG in recovering true covariance matrix. Further discussion is contained in Section 6.

2 SVAR and Its Identification

To describe the SVAR model we start with the usual K dimensional VAR model of order p (VAR(p)) represented as follows:

$$Y_t = a + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t \tag{1}$$

where $y_t = (y_{1t}, \ldots, y_{Kt})^T$ is a $(K \times 1)$ random vectors, the A_i are fixed $(K \times K)$ coefficient matrices, $a = (a_1, \ldots, a_K^T)$ is a $(K \times 1)$ vector of intercept. The K-dimensional white noise is given by $u_t = (u_{1t}, \ldots, u_{Kt}^T)$. In other words , we assume $E(u_t) = 0$, $E(u_t u_{ts}^T) = \Sigma_u$ for $t = s$ and $E(u_t u_s^T) = 0, t \neq s$.

A usual approach to find a model with contemporaneously uncorrelated residuals is to directly model the contemporaneous relations between the observed variables. It can be done as follows

$$\tilde{\mathsf{A}}Y_t = a + A_1^* y_{t-1} + \cdots + A_p^* y_{t-p} + \epsilon_t \tag{2}$$

, where $A_i^* = \tilde{\mathsf{A}}A_i$ $(i = 1, \ldots, p)$ and $\epsilon_t = \tilde{\mathsf{A}}u_t$, such that ϵ_t has a diagonal covariance matrix

$$\Sigma_\epsilon = \tilde{\mathsf{A}}\Sigma_u\tilde{\mathsf{A}}^T \tag{3}$$

Note that we need specific restrictions to make sure that matrix $\tilde{\mathsf{A}}$ is unique. From the relation (4) and the assumption that Σ_ϵ is diagonal matrix, we get $K(K-1)/2$ independent equation. In order for all K^2 elements of matrix $\tilde{\mathsf{A}}$ to have a unique solution we need $K(K+1)/2$ additional equations. By normalizing the diagonal elements of A to unity, we still need another $K(K-1)/2$ zero restrictions. If we assume a Wold causal ordering, then the matrix $\tilde{\mathsf{A}}$ is a lower-triangular matrix. Thus, we have exactly $K(K-1)/2$ restrictions and the associated impulse-responses are just-identified. Usually the literature uses just-identified SVAR and rely on theory to tell the recursion model. However if we believe that the SVAR is overidentified, we can use empirical based graphical models to represent causal relationships. In the next subsections we give overview of HVAR models.

2.1 Hierarchical Vector Autoregression

HVAR framework implements the lag order selection problem through convex regularization. The method prone to select low lag coefficients before corresponding high lag coefficients by forcing the selection to shrink toward low lag order solutions. In contrast to Song and Bickel (2011) approach, where they without enforcing a low-lag structure increase a weight of the penalty parameter with the coefficents' lag. For further convenience, as in Nicholson et al. (2016) we represent equation 1 in following form:

$$\begin{aligned}
&\mathbf{Y} := [\mathbf{y}_1 \ldots, \mathbf{y}_T] && (K \times T); && A := [A_1, \ldots, A_p] && (K \times Kp) \\
&\mathbf{Y}_t^0 := [\mathbf{y}_{t-1}^T, \ldots, \mathbf{y}_{t-p}^T]^T && (Kp \times 1); && \mathbf{X} := [\mathbf{Y}_1^0, \ldots, \mathbf{Y}_T^0] && (Kp \times T) \\
&\mathbf{U} := [\mathbf{u}_1, \ldots, \mathbf{u}_T] && (K \times T); && \mathbf{1} := [1, \ldots, 1]^T && (T \times 1)
\end{aligned}$$

Then, we can write equation1 as

$$\mathbf{Y} = \mathbf{a1}^T + \mathbf{AX} + \mathbf{U} \tag{4}$$

Using equation4 the traditional least squares can be expressed as minimizing

$$\min_{a,A} ||\mathbf{Y} - \mathbf{a1}^T - \mathbf{AX}||_2^2 \tag{5}$$

,where $||\cdot||_2$ is Frobenius norm. When number of observation T is not large compare to parameter space, the least squares results are unreliable. Nicholson et al. (2016) introduced three following different structures on the parameter space to make estimation possible.

- **Componentwise(C)**:This structure assumes that each of the K marginal equations in equation!1 has their own maximum lag, however each component within each equation has the same maximal lag.

- **Own-Other(O)**: This assumptions implies that a series own lag contains more information than other lags. Therefore, in this structure the diagonal elements are prioritized compare to off-digaonal elelments within each lag.

- **Elementwise(E)**: This structure is the most flexible one compare to other two. It does not assume any relationship among the coefficient lags.

In this paper, we consider only Componentwise structure, however our simulation results shows that in most of the times NPNDAG perform better than the usual approach as in Bessler and Akleman (1998) for the other two structural cases.

The convex optimization problem for Componentwise structural assumption after demeanizing data is following:

$$\min_A \left\{ \frac{1}{2} ||\mathbf{Y} - \mathbf{AX}||_2^2 + \lambda \sum_{i=1}^{K} \sum_{l=1}^{p} ||A^i_{(l:p)}||_2 \right\} \tag{6}$$

, where $A^i_{(l:p)} := [A^i_l \dots A^i_p]$ and A^i_l is the i'th row of the coefficient matrix at lag l. In equation 6, if the penalty $\lambda \geq 0$ is increasing than estimated $\hat{\mathbf{A}}^i_{(l:p)} = \mathbf{0}$ for more i and for smaller l. As well, if $\hat{A}^i_l = 0$, then $\hat{A}^i_h = 0$ for all $h > l$.

The modeling is done using so called hierarchical group lasso (Zhao et al., 2009), which is

modified version of a group lasso (Yuan and Lin, 2006) and allows a nested group structure. The group lasso penalty allows a set of groups of parameters to be simultaneously zero and nested group penalty assumes hierarchical sparsity constraints, that is if one set of parameters is zero than another set is also zero. The optimization problem in equation 4 can be efficiently solved using the proximal gradient method (Jenatton et al., 2011), which is extended version of gradient descent method, when the objective function is not smooth. The Algorithm 1 and 2 on page 9 at Nicholson et al. (2016) gives the general algorithm for solving HVAR problem.

3 Directed Acyclic Graphs and PC algorithm

A graph is a data structure $\mathcal{G}$ consisiting of a set of nodes and a set of edges. A pair of nodes X_i, X_j can be connected by a directed edge $X_i \to X_j$ or an undirected edge $X_i - X_j$. Thus, the set of edges ξ is a set of pairs, where each pair is one of $X_i \to X_j$,$X_i \leftarrow X_j$, or $X_i - X_j$. We say that $X_1 \dots X_k$ form a path in the graph $\mathcal{G}$ if, for every $i = 1, \dots k-1$, we have that either $X_i \to X_{i+1}$, or $X_i - X_{i+1}$. A path is directed, if, for at least one i, we have $X_i \to X_{i+1}$. A cycle in $\mathcal{G}$ is a directed path $X_1 \dots X_k$ where $X_1 = X_k$. A graph is acyclic if it contains no cycles. We called these graphs Directed Acyclic Graphs (DAG). DAGs are the fundamental graphical representation that underlies Bayesian Networks. A Bayesian Network Structure $\mathcal{G}$ is a directed acyclic graph whose nodes represent random variables $X_1 \dots X_n$. Denote $Pa^{\mathcal{G}}_{X_i}$ the parents of X_i in $\mathcal{G}$, and $NonDescendants_{X_i}$ the variables in the graph that are not descendents of X_i. Then $\mathcal{G}$ encodes following set of conditional independence assumptions, called the local independencies:

$$(X_i \perp NonDescendants_{X_i} | Pa^{\mathcal{G}}_{X_i}) \tag{7}$$

or

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i || Pa^{\mathcal{G}}_{X_i}) \tag{8}$$

The Independence implied by equation (7) and (8) can be read off the graph using the notion of d-separation (Pearl, 1986).Geiger et al. (1990) show the soundness and completeness of d-separation. By soundness they show that any independence reported by d-separation is satisfied by the underlying distribution. For completeness of d-separation, they need the notion of faithfulness. A

distribution is faithful to $\mathcal{G}$, if any independence in distribution is reflected in the d-separation properties of the graph. It can be shown that for almost all distributions that satisfy (8) over $\mathcal{G}$, that is, for all distributions except for a set of measure zero in the space of conditional probability distribution(CPD) parametrizations, faithfulness holds. In other words, for almost all possible choices of CPDs for the variables, the d-separation precisely characterizes the independence of the underlying distribution. (Koller and Friedman, 2009) The skeleton of a DAG $\mathcal{G}$ is obtained by substituting undirected edges for directed edges. Chickering (2002) showed that a probability distribution P, which is generated from a DAG $\mathcal{G}$ has a whole equivalence class of DAGs. Verma and Pearl (1991) characterizes the two DAGs in equivalent classes if and only if they have the same skeleton and the same $v-$structure. Where $v-$structure in a DAG $\mathcal{G}$ is a ordered triple of nodes (X_i, X_j, X_k) such that $X_1 \rightarrow X_2 \leftarrow X_3$.

In literature the representation of equivalent classes is done using the notion of partially directed acyclic graphs (PDAG), which is a graph where some edges are directed and some are undirected. a PDAG is completed (CPDAG), if every directed edge can be found alss in DAG's belonging to the same equivalence class and for every undirected edge $X_i - X_j$ one can find at least two DAGs in equicalent class with $X_i \rightarrow X_j$ and $X_i \leftarrow X_j$. Then, one can show that two CPDAGs are identical if and only if they represent the same equivalence class.

3.1 PC algorithm

PC algorithm is designed to estimate CPDAG. Usually CPDAG estimation consists of two parts: Skeleton estimation and partial orientation of edges. As usual in developing statistical estimation, there exist population and sample version of PC-algorithm. Algorithm1 and Algorithm 2 in Kalisch and Bühlmann (2007) represent a high dimensional consistent version of PC-algorithm to produce the correct skeleton and the CPDAG . In this population version of algorithm the assumption is that perfect knowledge about all necessary conditional independence relations is available. In general, the PC algorithm is an ordered set of commands that begins with a complete, undirected graph that places an undirected edge between every variable in the system (every variable in graph $\mathcal{G}$ vertex set $\mathcal{X}$) . Edges between variables are removed sequentially on the basis of zero correlations or zero partial (conditional) correlations. These conditioning variables on removed edges between

variables comprise the "sepset" of the variables whose edges has been removed. .The goal is to impose a directed edge among sets of variables X_1, X_2, X_3 in a vertex set (variable set)$\mathcal{X}$: $X_1 \to X_2 \to X_3, X_1 \leftarrow X_2 \leftarrow X_3, X_1 \to X_2 \leftarrow X_3$.

For finite sample version of PC-algorithm, the researcher need to estimate conditional independencies from the data. When the nodes of graph $\mathcal{G}$ has a multivariate normal distribution and the model is faithful, then from well-known property of the multivariate normal the partial correlations can be used as an estimates of conditional normal distribution(for proof see Proposition 2 in Kalisch and Bühlmann (2007) or Proposition 5.2 in Lauritzen (1996)). That is

$$X_u \perp X_v | X_S \iff \rho_{uv|S} = 0, \tag{9}$$

, where the partial correlation $\rho_{uv|S}$ for any $w \in S$ is given by

$$\rho_{uv|S} = \frac{\rho_{uv|S\setminus w} - \rho_{uv|S\setminus w}\rho_{vw|S\setminus w}}{\sqrt{(1-\rho^2_{uw|S\setminus w})(1-\rho^2_{vw|S\setminus w})}} \tag{10}$$

4 PC algorithm for Gaussian Copula based Graphical Models

In this section we give overview of high-dimensional consistency properties of PC-algorithm (Harris and Drton, 2013) for a broader class continuous distributions with Gaussian Copulas or as named in Liu et al. (2009) nonparanormal distributions. In this algorithm, to test the conditional independence the Pearson-type sample correlations are replaced by rank-based measures of correlations such as Kendall's τ and Spearman's ρ. We start by defining the nonparanormal distribution.

Definition 1. A random vector $X = [X_1 \dots X_p]^T$ has a nonparanormal distribution if there exist functions $\{f_j\}_{j=1}^p$ such that $Z := f(\mathbf{X}) \sim N(0, \Sigma)$, where $f(\mathbf{X}) = [f_1(X_1), \dots, f_p(X_p)]^T$.

Considering f_v functions as affine, then by definition, all multivariate normal distributions are also nonparanormal. Let $X \sim NPN(f, \Sigma)$, then the marginal distribution for specific X_j coordinate may have CDF F_j then:

$$F_j(x) = P(X_j \le x) = P(Z_j \le f_j(x)) = \Phi\Big(\frac{f_j(x) - \mu_j}{\sigma_j}\Big)$$

which implies that

$$f_j(x) = \sigma_j \Phi^{-1}(F_j(x))$$

Lemma 1. *The nonparanormal distribution $NPN(\Sigma, f)$ is a Gaussian copula when the f_j's are monotone and differentiable.*

The proof of the lemma is the simple use of Sklar's theorem (Nelsen, 2006) and can be found in Liu et al. (2009, page 2298).

Definition 2. The nonparanormal graphical model $NPN(G)$ associated with a DAG G is the set of all distributions $NPN(f, \Sigma)$ that are Markov[2] with respect to $\mathcal{G}$.

From the deterministic chatacterization of marginal transformations f_v, the dependence structure in a nonparanormal distribution corresponds to the underlying latent multivariate normal distribution. That is , if $X \sim NPN(f, \Sigma)$ and $Z \sim N(0, \Sigma)$, then it is true that for any triple of pairwise disjoint sets $A, B, S \subset V$

$$X_A \perp X_B | X_S \iff Z_A \perp Z_B | Z_S$$

Therefore, for any two nodes u and v and a seperating set $S \subset V \backslash \{u, v\}$, it is true that

$$X_u \perp X_v | X_S \iff \rho_{uv|S} = 0, \tag{11}$$

, where the partial correlation $\rho_{uv|S}$ for any $w \in S$ is given as in Equation 10

Based on equivalence (11) Harris and Drton (2013) concluded that

$$X_u \perp X_v | X_S \iff \hat{\rho_{uv|S}} \leq \gamma \tag{12}$$

, where $\rho_{uv|S}$ is rank-based correlation estimate and $\gamma \in [0, 1]$ is a fixed threshold. They refer PC algorithm that uses the conditional independence tests from (12) as the 'Rank PC' (RPC) and demonstrate the high-dimensional consistency of RPC. For details, see Harris and Drton (2013). We use RPC algorithm as a fundamental block to build the NPNDAG procedure.

[2]The distribution is Markov if it satisfies equations 7 and 8

4.1 Directed Graphs and SVAR

The usual procedure to use DAG's in SVAR is as follows: treat the estimated innovations from equation (2) as the original data (Swanson and Granger, 1997; Bessler and Akleman, 1998; Demiralp et al., 2014). Then the estimated covariance matrix is considered as an input for PC algorithm to compute the various conditional correlations. The output of the algorithm with corresponding zeros corresponds to particular zeros in $\tilde{\mathsf{A}}$ in (2). Often times PC algorithm returns only partially directed acyclic graphs, is such is the case one can use bootstrap methods, as described in Demiralp et al. (2014) to refine the restrictions. The NPNDAG procedure can be described as follows:

Algorithm 1 NPNDAG procedure

procedure NPNDAG
input:
    $\Sigma \leftarrow K \times K$ *correlation matrix of residuals*
    $\alpha \leftarrow$ Confidence level for *conditional independence test*
top:
    Run *RPC algorithm*
    Obtain *Contemporaneous time restrictions for* $\tilde{A}$
    Maximize *Likelihood* $C + \frac{T}{2} ln|\tilde{A}|^2 - \frac{T}{2} tr(\tilde{A}^T \tilde{A} \hat{\Sigma}_u)$
Output:
    $\tilde{A}$

,where $\hat{\Sigma_u}$ is estimated residuals and $\tilde{A}$ as in equation 2. Usually, the maximization problem does not have closed form solution and optimization is done numerically using gradient descent algorithm. More details can be found in Lutkepohl (2007, Section 9.3) .

5 Numerical Results

We analyze the NPNDAG procedure for finding the contemporaneous correlation matrix using various simulated data sets and macroeconomic data. The numerical results have been obtained using software $\mathbb{R}$ .For PC (RPC)-algorithm we use $\mathbb{R}$-package pcalg (Kalisch et al., 2012) and for Lasso-VAR package BIGVAR (Nicholson et al., 2017).

5.1 Simulating Data

In order to simulate stationary time series, we first generate covariance matrix Σ_u by constructing an adjacency matrix $\mathbf{B}$ as described in Kalisch and Bühlmann (2007):

- Fix an ordering of the variables.
- Fill the adjacency matrix $\mathbf{B}$ with zeros.
- Replace every matrix entry in the lower triangle (below the diagonal) by independent realizations of Bernoulli(s) random variables with success probability s where $0 < s < 1$. We will call s the sparseness of the model.
- Replace each entry with a 1 in the adjacency matrix by independent realizations of a *Uniform*[0.1,1] random variable.

The mentioned steps generate the matrix B whose entries are zero or in the range [0.1,1]. The corresponding DAG draws a directed edge from node i to node j if $i < j$ and $B_{ji} \neq 0$. The DAGs that are created in this way have the following property: $\mathbb{E}[N_i] = s(p-1)$, where N_i is the number of neighbors of a node i. Therefore, when sparseness parameter s is low the DAG has few neighbors and vice-versa. In order to construct a stationary coefficient matrix A for VAR process, we start by converting the $\mathbf{VAR}_k(p)$ to $\mathbf{VAR}_k(1)$ as described in equation 2.1.8 of Lutkepohl (2007).Than we enforce its maximum eigenvalue be less than 1 to assure the stationarity Lutkepohl (2007). Thus

- Generate u_t for the two distributions , we consider $N(0, \Sigma_u)$ and $t_3(0, \Sigma)$.
- Generate sparse random matrix for $A_1, \ldots, A_p$.
- Convert $\mathbf{VAR}_k(p)$ to $\mathbf{VAR}_k(1)$
- Generate Y corresponding to equation 1.

After generating the data Y we run HVAR model and obtain residuals for further analysis.

5.1.1 Evaluate Performance for Different Parameter Setting

To assess the quality of fit we follow an approach suggested by Tsamardinos et al. (2006); Kalisch and Bühlmann (2007) and use Structural Hamming Distance (SHD). SHD counts the number of edge insertion, deletions, and flips in order to transfer the PC output into the correct DAG. Therefore, a large SHD indicates a poor fit and vice versa. We fit 100 replications to all combinations of

- $p \in \{10, 40\}$
- $n \in \{250, 500, 1000\}$
- $s \in \{0.1, 0.4\}$
- $\alpha = 0.1$

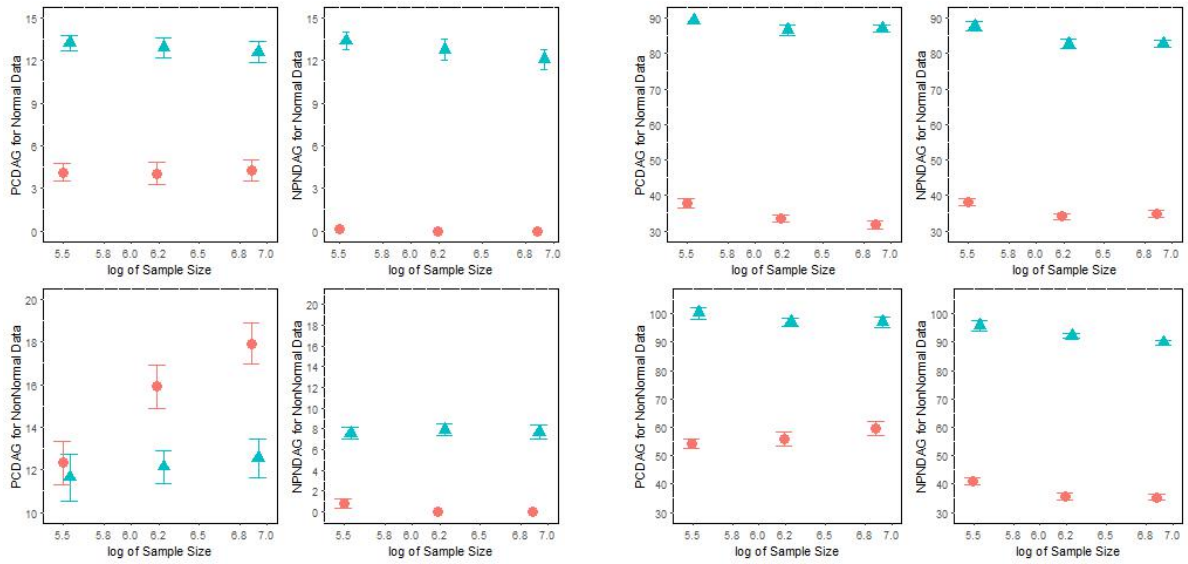
5.1.2 Results

Figure 1 reports output from PCDAG and NPNDAG algorithms using settings defined on Section 5.1.1. We take the logarithm of sample size to make the representation simpler. $\diamond$ and $\triangle$ corresponds to sparsity level $s \in \{0.1, 0.4\}$ respectively. Figure and corresponds to number of variables $p = 10$ and $p = 40$, respectivelly.The upper row is the output for Normal data and the bottom row for nonnormal $\mathbf{t}$ distribution.

We can see that in case of $p = 10$ NPNDAG algorithm outperform the PCDAG algorithm for both Normal and non normal data as well for two sparsity levels. In case of $p = 40$ NPNDAG algorithm doing as well as PCDAG when data is generated from Normal distribution and outperforms PCDAG for non-normal case.

5.2 Macroeconomic Dataset

We show the application of NPNDAG on a real-world macroeconomic dataset. The data set represents the 168 monthly US macroeconomic time series from 01/1959 to 02/2009. Initially the dataset was compiled by Stock and W. Watson (2005) and augmented by Koop (2011). The full list of variables can be found in Koop (2011). In this paper we consider only *Medium-Large* data



(a) $p = 10$ (b) $p = 30$

Figure 1: $\diamond$ and $\triangle$ corresponds to sparsity level $s \in \{0.1, 0.4\}$ respectively. a)Output from PCDAG and NPNDAG algorithms for $p = 10$ for the sample size$n \in \{250, 500, 1000\}$ and $\alpha = 0.1$ b)Output from PCDAG and NPNDAG algorithms for $p = 30$ for the sample size$n \in \{250, 500, 1000\}$ and $\alpha = 0.1$.

set ($k = 40$). More information about the *Medium-Large* data set and as well for other types of data-sets can be found in Koop (2011).

First we transform the data-set to make the variables approximately stationary.[3] Then we standardize each series to have mean 0 and variance 1 as recommended in Nicholson et al. (2016) and run HVAR. Before imposing structure on residuals obtained from HVAR we estimate and plot kurtosis and skewness for the data-set to visually verify the normality assumption. Figure 2 indicates that the normality assumption for residuals is not valid. The skeweness is around -2 for more than 30 variables then it grows exponentially and the logarithm of kurtosis starts around 2(or around 8 without transformation) and has monotonic increasing behavior.

We estimate the contemporaneous time correlation for 40 residuals using PCDAG as in Bessler and Akleman (1998) and NPNDAG at significance level for the individual conditional independence tests at $\alpha = 0.05$. Figures (3 and 4) show the output of RPC and PC algorithms, respectively .The NPNDAG produces total 2 undirected edges and 60 directed. The maximum number of neighbors is 3 and the average number of neighbors is around 1.6. For the PCDAG the number of undirected

[3]Koop (2011) gives the detailed about the transformation.

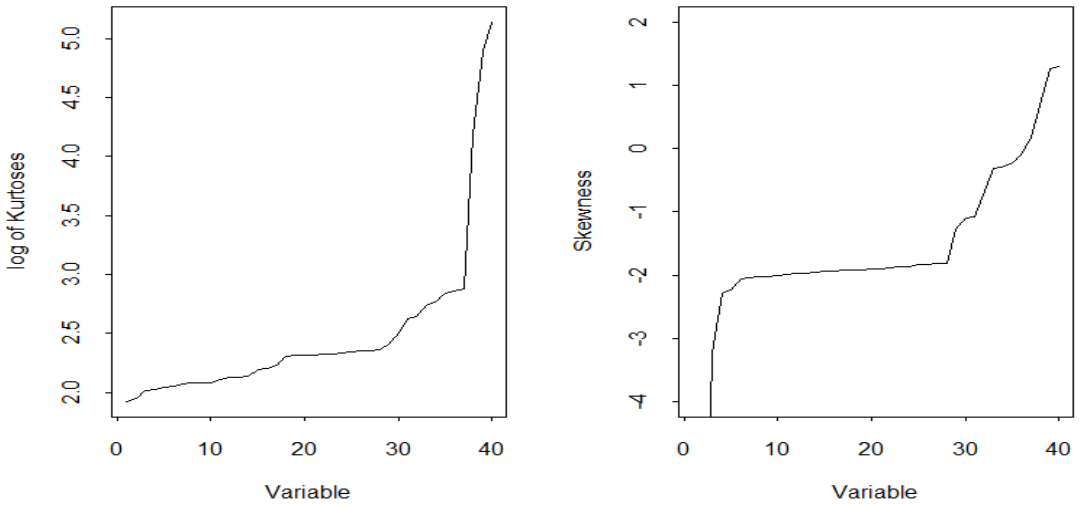


Figure 2: Plot of Logarithm of Kurtosis and Skewness for Large Dataset

edges is 7 and the number of directed edges is 43 . The Structural Hamming Distance between two graphs is 56, which indicates the difference of two methods.

Using the contemporaneous time restrictions produced by two algorithms, we are ready to find the impulse response functions (IRF). Figure 5 represents plots of response of Consume Price Index to the first 20 variables in *Medium-Large* model. Results show that for some shocks the response produced from the two algorithms are identical, however for most of the shocks there is either difference in magnitude of shock or the outputs significantly different from each other. In this paper, we do not interpret or give meaning to results, since our main task is different.

6 Concluding Remarks

Disregarding the contemporaneous time correlation among the residuals may effect the interpretation of IRF and FEV. In this paper we suggest the extended version of PCDAG algorithm to account for non-Nonnormal distribution using the latest advance techniques in Machine Learning and Statistical Learning literature. Our simulation results suggested that NPNDAG was able to perform as well as the PCDAG in case of Normal distribution and outperform PCDAG for non-Normal distribution.

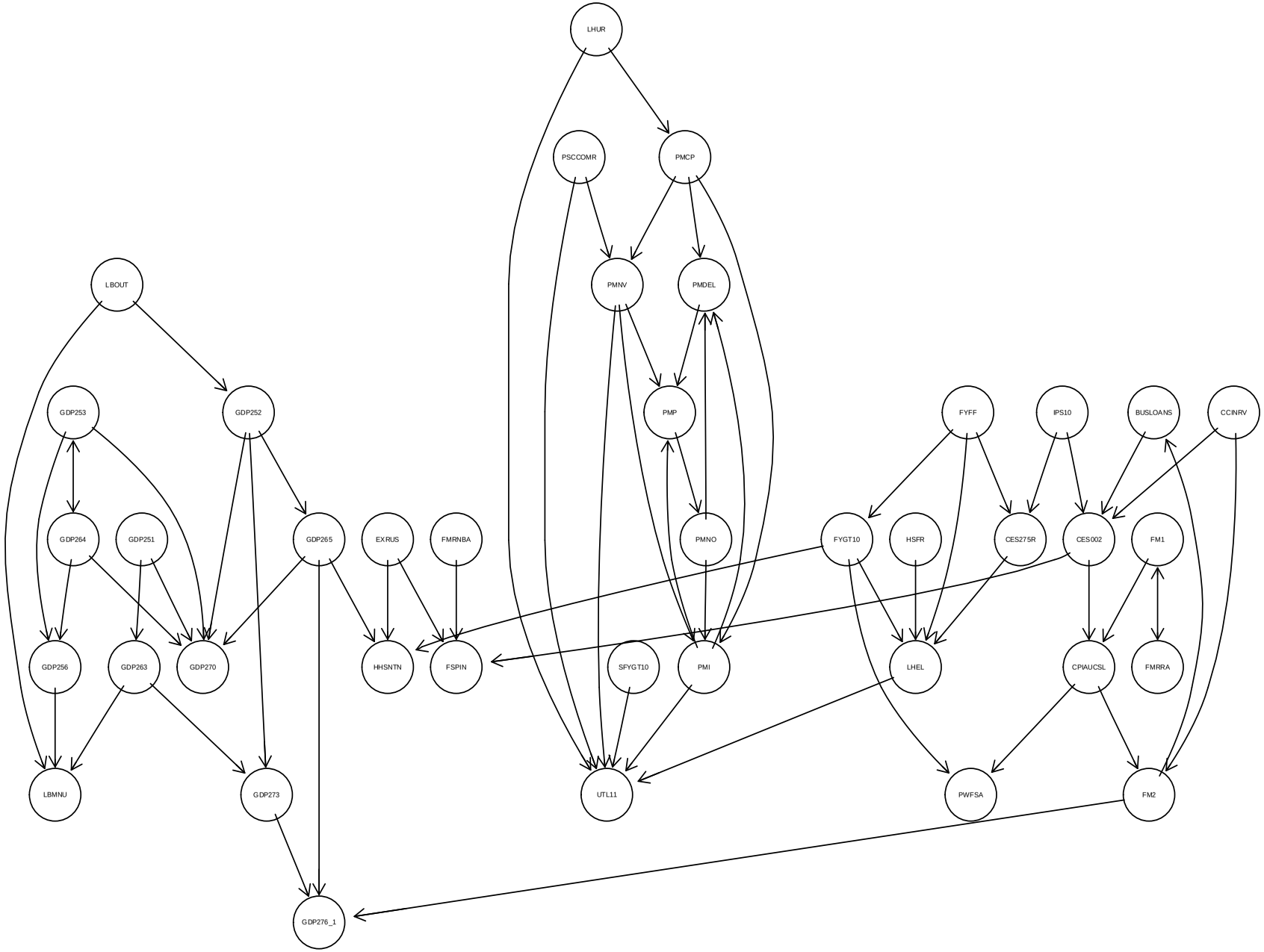


Figure 3: Output of RPC Algorithm

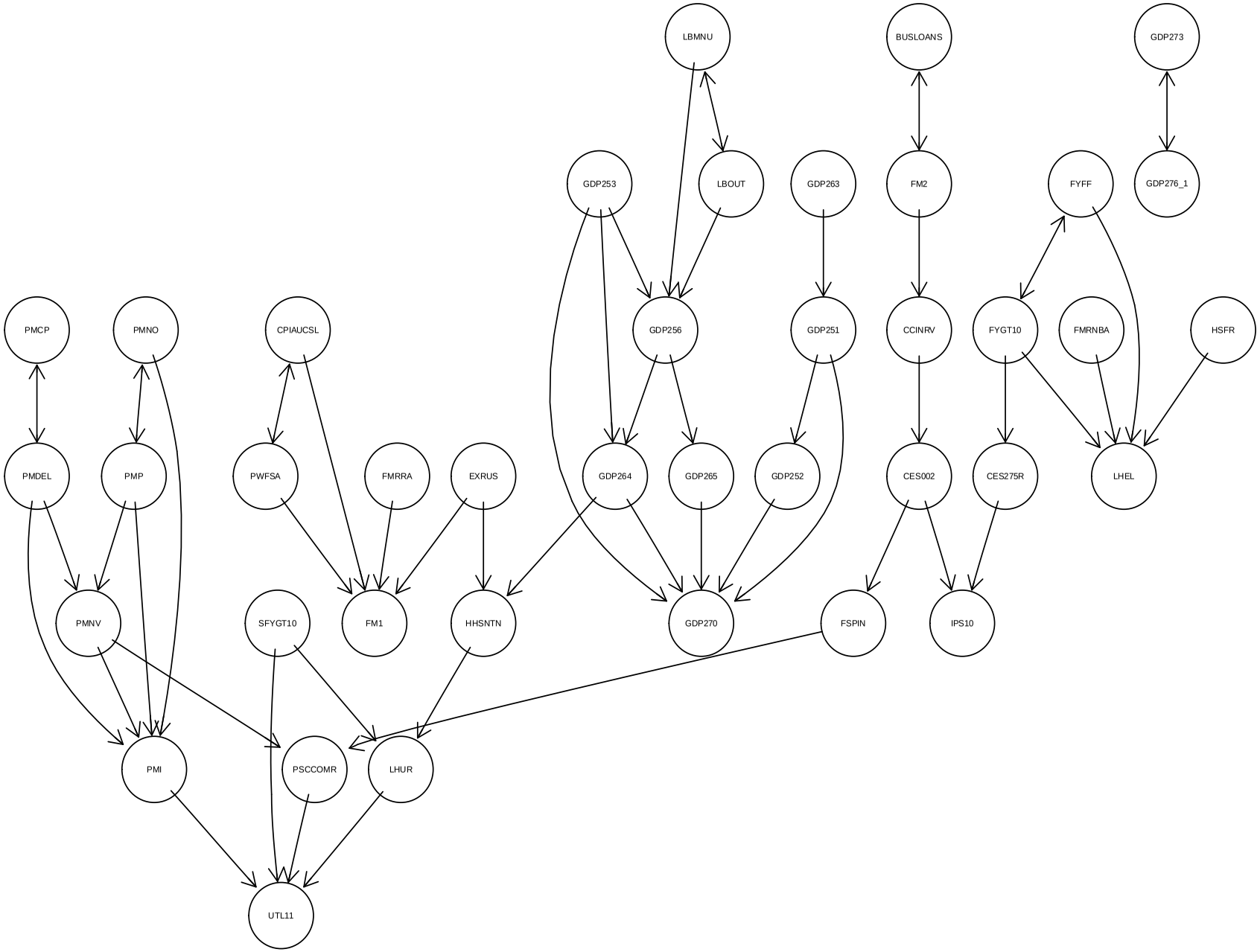


Figure 4: Outpur of PC algorithm

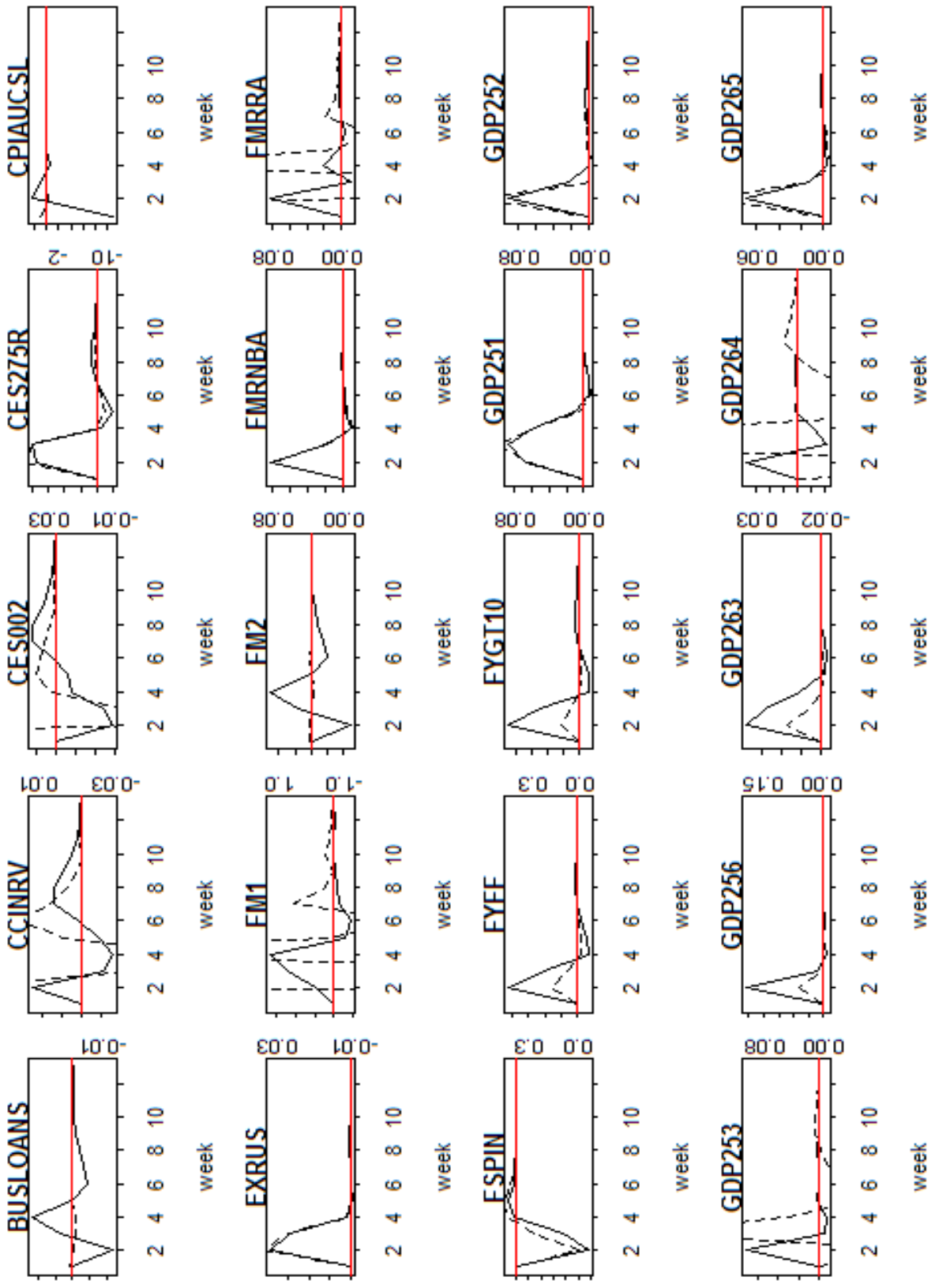


Figure 5: Impulse response functions(IRF) for model *Medium-Large*. The figure presents the response of Consumer Price Index to shocks of the first 20 variables in *Medium-Large* model.

References

Banbura, M., D. Giannone, and L. Reichlin (2010), "Large bayesian vars." *Journal of Applied Econometrics*, 25, 71–92.

Bernanke, B. (1986), "Alternative explanations of the money-income correlation." *Carnegie-Rochester Conference Series on Public Policy.*, 45–100.

Bessler, B. (1984), "An analysis of dynamic economic relationships: An application th the u.s hog market." *Canadian Journal of Agricultural Economics*, 109–24.

Bessler, D. and D. Akleman (1998), "Farm prices,retail prices, and directed graphs: Results for pork and beef." *American Journal of Agricultural Economics.*, 1144–1149.

Bessler, David A. and Seongpyo Lee (2002), "Money and prices: U.s. data 1869–1914." *Empirical Economics*, 27, 427–446.

Chickering, David Maxwell (2002), "Learning equivalence classes of bayesian-network structures." *J. Mach. Learn. Res.*, 2, 445–498.

Davis, Richard A., Pengfei Zang, and Tian Zheng (2016), "Sparse vector autoregressive modeling." *Journal of Computational and Graphical Statistics*, 25, 1077–1096.

Demiralp, Selva, Kevin Hoover, and Stephen Perez (2014), "Still puzzling: evaluating the price puzzle in an empirically identified structural vector autoregression." *Empirical Economics*, 46, 701–731.

Demiralp, Selva and Kevin D. Hoover (2003), "Searching for the causal structure of a vector autoregression*." *Oxford Bulletin of Economics and Statistics*, 65, 745–767.

Doan, T., R. Litterman, and C. A. Sims (1984), "Forecasting and conditional projection using realistic prior distributions." *Econometric Reviews*, 3, 1–100.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), "The generalized dynamic factor model: identification and estimation." *Review of Economics and Statistics*, 82, 540554.

Geiger, Dan, Thomas Verma, and Judea Pearl (1990), "Identifying independence in bayesian networks." *Networks*, 20, 507–534.

Haigh, M. and D. Bessler (2004), "Causality and price discovery: An application of directed acyclic graphs." *Journal of Business.*, 1097–1121.

Harris, Naftali and Mathias Drton (2013), "Pc algorithm for nonparanormal graphical models." *J. Mach. Learn. Res.*, 3365–3383.

Jenatton, Rodolphe, Julien Mairal, Guillaume Obozinski, and Francis Bach (2011), "Proximal methods for hierarchical sparse coding." *J. Mach. Learn. Res.*, 12, 2297–2334.

Kalisch, Markus and Peter Bühlmann (2007), "Estimating high-dimensional directed acyclic graphs with the pc-algorithm." *J. Mach. Learn. Res.*, 8, 613–636.

Kalisch, Markus, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann (2012), "Causal inference using graphical models with the R package pcalg." *Journal of Statistical Software*, 47, 1–26, URL `http://www.jstatsoft.org/v47/i11/`.

Koller, Daphne and Nir Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Koop, Gary M (2011), "Forecasting with medium and large bayesian vars." *Journal of Applied Econometrics*, 28, 177–203.

Lauritzen, Steffen L. (1996), *Graphical Models*. Oxford University Press.

Litterman, R. (1986), "Forecasting with bayesian vector Autoregressions Five Years of Experience." *Journal of Business and Economic Statistics*, 4, 2538.

Liu, H., J. Lafferty, and L. Wasserman (2009), "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *J. Mach. Learn. Res.*, 2295–2328.

Liu, Han, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman (2012), "High-dimensional semiparametric gaussian copula graphical models." *Ann. Statist.*, 40, 2293–2326.

Lutkepohl, H. (2007), *New Introduction to Multiple Time Series Analysis*. New York:Springer.

Nelsen, Roger (2006), *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Nicholson, W., D. Matteson, and J. Bien (2017), "BigVAR: Tools for Modeling Sparse High-Dimensional Multivariate Time Series." *ArXiv e-prints*.

Nicholson, William B., Jacob Bien, and David S. Matteson (2016), "Hierarchical vector autoregression." *Arxiv preprint arXiv:1412.5250v2*.

Pearl, J. (1986), "Fusion, propagation and structuring in belief networks." *Artificial Intelligence*, 241–288.

Pearl, Judea (2009), *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, New York, NY, USA.

Shojaie, A. and G. Michailidis (2010), "Discovering graphical granger causality using the truncating lasso penalty." *Bioinformatics*, 26, 517–523.

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

Song, S. and P. J. Bickel (2011), "Large vector auto regressions." *Arxiv preprint arXiv:1106.3915*.

Spirtes, P., C. Glymour, and R. Scheines (2000), *Causation, Prediction, and Search*, 2nd edition. MIT press.

Stock, J. H. and M. W. Watson (2002), "Forecasting using principal components from a large number of predictors." *Journal of the American Statistical Association*, 97, 147162.

Stock, James and Mark W. Watson (2005), "An empirical comparison of methods for forecasting using many predictors." *Manuscript, Princton University*.

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

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasoo." *Journal of the Royal Statistical Society*, 58, 267–288.

Tsamardinos, Ioannis, Laura E. Brown, and Constantin F. Aliferis (2006), "The max-min hill-climbing bayesian network structure learning algorithm." *Mach. Learn.*, 65, 31–78.

Verma, Thomas and Judea Pearl (1991), "Equivalence and synthesis of causal models." In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, 255–270, Elsevier Science Inc., New York, NY, USA.

Yuan, M. and Y. Lin (2006), "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society, Series B*, 68, 49–67.

Zhao, P. and B Yu (2006), "On model selection consistency of lasso." *Journal of Machine Learning Research*, 7, 2541–2563.

Zhao, Peng, Guilherme Rocha, and Bin Yu (2009), "The composite absolute penalties family for grouped and hierarchical variable selection." *Ann. Statist.*, 37, 3468–3497.