Bayesian Graphical Models for Structural Vector Autoregressive Processes

Daniel Ahelegbey, Monica Billio, and Roberto Cassin (2014)

Presented by: Jacob Warren

March 21, 2015

Background

- This paper is about combining information from Dynamic Networks to inform the causal structure of Structural Vector Autoregressions
- The paper discusses using networks for estimating both structural and autoregressive coefficients, but my focus will be on the structural components (the autoregressive ones are easier)

Background: What is a SVAR?

- Consider the structural VAR(L) process:

$$Y_t = B_0 Y_t + B_1 Y_{t-1} + \cdots + B_L Y_{t-L} + \varepsilon_t$$

Where $\varepsilon_t \sim N(0, I_p)$, and B_0 has 0s on the diagonal

Background: What is a SVAR?

- Consider the structural VAR(L) process:

$$Y_t = B_0 Y_t + B_1 Y_{t-1} + \cdots + B_L Y_{t-L} + \varepsilon_t$$

  Where $\varepsilon_t \sim N(0, I_p)$, and B_0 has 0s on the diagonal

- Rewriting the SVAR into a reduced-form VAR:

$$Y_t = A_0^{-1} B_1 Y_{t-1} + A_0^{-1} B_2 Y_{t-2} + \ldots A_0^{-1} B_L Y_{t-L} + A_0^{-1} \varepsilon_t$$

  where $A_0 = \mathbf{I} - B_0$

- The problem is that the structural parameter, A_0 is not identified.

Background: Identification Issue

$$Y_t = A_0^{-1} B_1 Y_{t-1} + A_0^{-1} B_2 Y_{t-2} + \dots A_0^{-1} B_L Y_{t-L} + A_0^{-1} \varepsilon_t$$

- Observe that $A_0^{-1} A_0^{-1'}$ is the covariance of the error term.

Background: Identification Issue

$$Y_t = A_0^{-1} B_1 Y_{t-1} + A_0^{-1} B_2 Y_{t-2} + \dots A_0^{-1} B_L Y_{t-L} + A_0^{-1} \varepsilon_t$$

- Observe that $A_0^{-1} A_0^{-1'}$ is the covariance of the error term.
- But, for any orthogonal matrix Q (so $QQ' = I$),

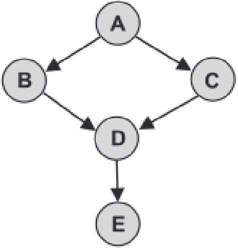
$$A_0^{-1} A_0^{-1'} = A_0^{-1} Q Q' A_0^{-1'} = (A_0^{-1} Q)(A_0 Q)'$$

- So the structural parameters, A_0 are not identified
  - Requires making some identification assumptions to perform structural analysis

Background: Directed Acyclic Graphs

- A *Directed Acyclic Graph* is a collection $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges.
  For example:

Figure 1 : DAG from Grzegorcyzk (2001)



- In this graph, node A is the parent of B and C. B and C are child nodes of A. D is the child node of both B and C, and E is the child node of E.
- A graph is said to be acyclic if no node is a descendant of itself

Connecting DAG and SVAR

$$Y_t = B_0 Y_t + B_1 Y_{t-1} + \cdots + B_L Y_{t-L} + \varepsilon_t$$

- Then, there is a one-to-one relationship between the regression matrices and Directed Acyclic Graphs, given as:

$$X^j_{t-s} \to X^i_t \iff B_s(i,j) \neq 0$$

Where $X^j_{t-s} \to X^i_t$ means that X^j_{t-s} "causes" in some way, X^i_t.

Connecting DAG and SVAR

$$Y_t = B_0 Y_t + B_1 Y_{t-1} + \cdots + B_L Y_{t-L} + \varepsilon_t$$

- Then, there is a one-to-one relationship between the regression matrices and Directed Acyclic Graphs, given as:

$$X^j_{t-s} \to X^i_t \iff B_s(i,j) \neq 0$$

Where $X^j_{t-s} \to X^i_t$ means that X^j_{t-s} "causes" in some way, X^i_t.

  - Note that despite the directed aspect, the assumption of causality is not totally innocuous
  - Think about a stock market ascent that may lead to an increase in GDP. The stock market may not *cause* GDP to increase, it could merely lead it in time.
  - For more information, see Dawid 2008

- This is similar to the notion of Granger Causality (with the exception of contemporaneous causation and conditional independence)

DAG and Cholesky

- One identification of the orthogonal shocks in the VAR is to use a Cholesky Decomposition

$$PP' = A_0^{-1} A_0^{-1'}$$

- Acyclicality implies a specific ordering of variables
- For example, if $X_1 \rightarrow X_2$, then let X_1 be the first variable in the system, and X_2 the second

Local Markov Property

- A graph is said to be posses the Local Markov Property if

$$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{N} P(X_i | pa(X_i))$$

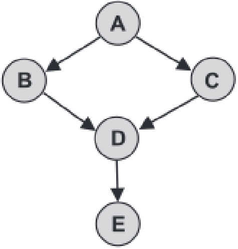
where $pa(X_i)$ is the set of parent nodes for node X_i

Local Markov Property

- A graph is said to be posses the Local Markov Property if

$$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{N} P(X_i | pa(X_i))$$

where $pa(X_i)$ is the set of parent nodes for node X_i



- In the example above, the full likelihood of the graph can be simplified into:

$$P(G) = P(A)P(B|A)P(C|A)P(D|\{B, C\})P(E|D)$$

Estimation

- Define $B_s^\star = G_s \circ \Phi_s, 0 \leq s \leq p$, where $\circ$ is the Hadamarad element-wise matrix multiplication, and Φ_s are the reduced form parameters
- G_s are connectivity matrices that indicates dependence

Estimation

- Define $B_s^\star = G_s \circ \Phi_s, 0 \leq s \leq p$, where $\circ$ is the Hadamarad element-wise matrix multiplication, and Φ_s are the reduced form parameters
- G_s are connectivity matrices that indicates dependence
- Thus, the reduced form parameters of the VAR can be written as:

$$A_0 = I - G_0 \circ \Phi_0$$
$$A_i = (I - G_0 \circ \Phi_0)^{-1}(G_i \circ \Phi_i)$$

Estimation: Bayesian Paradigm (priors)

$$Y_t = B_0^\star Y_t + B_1^\star Y_{t-1} + \cdots + B_L^\star Y_{t-L} + \varepsilon_t, \quad \varepsilon_t \sim N(0, I)$$

- Likelihood:
  - The data matrix, $\mathcal{X} \sim N(0, \Sigma_x)$
- Prior:
  - Define the probability distribution over a graph as:

    $$P(\mathcal{G}, \Theta) = P(\mathcal{G})P(\Theta|\mathcal{G})$$

    Where $\mathcal{G}$ is the set of graph structures (nodes, edges and directions), and Θ is the set of parameters.
  - $P(G) \propto 1$
  - B_i are distributed normally
  - Conditional on a complete graph, $P(\Sigma|G) \sim IW$

Estimation: Bayesian Paradigm (priors)

$$Y_t = B_0^\star Y_t + B_1^\star Y_{t-1} + \cdots + B_L^\star Y_{t-L} + \varepsilon_t, \quad \varepsilon_t \sim N(0, I)$$

- Likelihood:
  - The data matrix, $\mathcal{X} \sim N(0, \Sigma_x)$
- Prior:
  - Define the probability distribution over a graph as:

    $$P(\mathcal{G}, \Theta) = P(\mathcal{G})P(\Theta|\mathcal{G})$$

    Where $\mathcal{G}$ is the set of graph structures (nodes, edges and directions), and Θ is the set of parameters.
  - $P(G) \propto 1$
  - B_i are distributed normally
  - Conditional on a complete graph, $P(\Sigma|G) \sim IW$

Note: This seems a little strange, since they have not specified a prior on the covariance *except* conditional on a complete graph. If the graph is not complete, the covariance will not be IW distributed

Marginal Likelihood

- The marginal likelihood can be factorized

$$P(\mathcal{X}|G) = \int P(\mathcal{X}|G, \Sigma_x)P(\Sigma_x|G)d\Sigma_x$$

- Under the assumed likelihood/priors, the marginal likelihood has a closed form
- They estimate a Multivariate-Normal-Inverse-Wishart process and a Minnesota Prior process

Model Inference

- Since everything has closed forms, they implement the following Gibbs Sampler:
    1. Sample the graph from the conditional posterior using Metropolis Hastings
    2. Sample the reduced form parameters directly from their posterior

Simulations

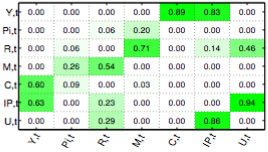
- They compare their scheme to a competitor (the PC algorithm)
- They find that their model does about 10% better for a small-scale VAR (n=5), but comparably well in a larger system (n=10)

Simulations

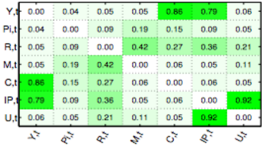
- They compare their scheme to a competitor (the PC algorithm)
- They find that their model does about 10% better for a small-scale VAR (n=5), but comparably well in a larger system (n=10)
- An additional advantage of their model over the PC algorithm is that they can do prediction

Macroeconomic Time Series application

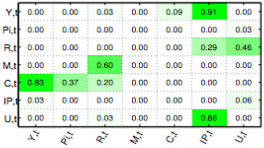
- Macro-forecasting based on medium size (n=20) VARs
- Dataset is quarterly observations from 1959Q1-2008Q4
- Rolling windows of 14 years are used to estimate the model



(a) PC

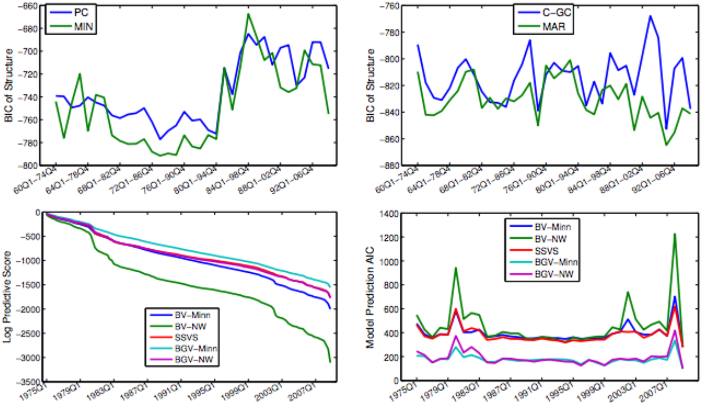


(c) SSVS(ω)



(e) MIN

Goodness of Fit



"Big" Data Implications

- It is hard to tell, but it seems like this estimation is extremely computationally intensive
  - Calculating the contemporaneous dependencies are limited to 5-7 variables
  - Time series are limited to to $\sim$ 50 observations
- They also apply their process to 19 financial sectors to estimate financial interconnectedness, but only estimate the autoregressive component

Comments

- This paper was overall very good
- Concise, Bayesian method for estimating structural parameters of an SVAR
- However, they do not tie their results into cholesky factorization or SVAR identification at the end
  - Should we do impulse responses with their method or others?
- Can the efficiency of the inference scheme be improved