

Identifiability of VAR(1) model in a stationary setting

Bixuan Liu

Supervised by: Stéphane Robin and Viet Chi Tran

Aussois, June 19th, 2025



Overview

- 1 Introduction
- 2 Identifiability framework in algebraic statistics
- 3 Identifiability Results

Introduction

VAR(1) in a stationary setting

The First-order Vector Autoregressive (VAR(1)) model:

$$\begin{cases} \mathbf{x}_1 = \epsilon_1, \\ \mathbf{x}_t = \Lambda^T \mathbf{x}_{t-1} + \epsilon_t \quad \text{for } t > 1, \\ \epsilon_t \sim \mathcal{N}(0, \Omega) \end{cases}$$

where $\Lambda \in \mathbb{R}^{n \times n}$ is a deterministic matrix. Assume $\Omega = \omega I_n$. If the eigenvalues of Λ are all smaller than 1 in absolute value, then as t goes to infinity, $\mathbf{x}_t$ converges in distribution to

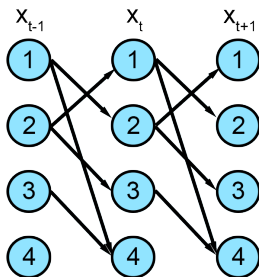
$$\mathbf{x}_\infty \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{n \times n}$ satisfies

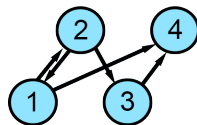
$$\Sigma = \Lambda^T \Sigma \Lambda + \omega I_n.$$

Assume Σ is known, can we recover the support of Λ , i.e. the underlying graph structure G ?

Graphical model



(a) The graphical VAR(1) model



(b) The encoded graph G

Motivation

Why VAR(1)?

The fluctuation process around the solution of the Lotka-Volterra (LV) equations¹:

$$\mathbf{x}_t = \sqrt{n}(\boldsymbol{\eta}_t^n - \boldsymbol{\eta}_t^*)$$

- When $n \rightarrow +\infty$, $\mathbf{x}_t$ converges to a **centered Gaussian process**.

Robert May's model²:

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{\Lambda}^T \mathbf{x}_t.$$

- VAR(1) is simply a **random version** of May's model.
- VAR(1) is suitable for a subject **rarely explored**.

1. Akjouj, Imane, et al. "Complex systems in ecology: a guided tour with large Lotka–Volterra models and random matrices." Proceedings of the Royal Society A 480.2285 (2024): 20230284.

2. May, R. M. Will a large complex system be stable? Nature 238, 5364 (1972), 413–414.

Motivation

Why the stationary regime? / Why not time-series analysis?

Central topic: reconstruct the dynamic interaction network from the abundance data captured in the ecosystem.

Sampling insufficiency problem¹

Time-series sampling is **too costly** and are often sacrificed in practice. Instead, the limited resources are used to cover as much ground as possible.

Example: Study of a global ocean cross-domain plankton co-occurrence network with one-time-samplings of 200 stations².

- Results from time-series analysis are **inapplicable**.

1. Legendre, Pierre, Miquel De Cáceres, and Daniel Borcard. "Community surveys through space and time: testing the space-time interaction in the absence of replication." *Ecology* 91.1 (2010): 262-272.

2. Chaffron, S., et al. "Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Sci Adv.* 7 (35): eabg1921." 2021.

Recall

$$\Sigma = \Lambda^T \Sigma \Lambda + \omega I_n.$$

Majority of previous work $\rightarrow$ **time series**.

Young's condition¹ - counting the parameters

A necessary condition for identifiability is that the number of non-zero off-diagonal elements of Λ be no more than $n(n-3)/2$ and that $n \geq 5$.

Drawback of Young's condition:

- Too weak;
- Lack of interpretability;
- Not a sufficient condition.

1. Young, William Chad, Ka Yee Yeung, and Adrian E. Raftery. "Identifying dynamical time series model parameters from equilibrium samples, with application to gene regulatory networks." Statistical modelling 19.4 (2019): 444-465.

Identifiability framework in algebraic statistics

Parametrization

$G = (V, \mathfrak{E}_G)$: the graph corresponding to the support of Λ of a VAR(1) model.

Parameter space

Define:

$$M_G := \{\Lambda = (\lambda_{ij}) \in M_n(\mathbb{R}) \mid \rho(\Lambda) < 1, \text{ and } \lambda_{ij} = 0 \text{ if } (i, j) \notin \mathfrak{E}_G\},$$

where $\rho(\cdot)$ is the spectral radius. The **parameter space** is then $M_G \times \mathbb{R}^+$.

Parametrization map

$$\phi_G : M_G \times \mathbb{R}^+ \rightarrow M_n(\mathbb{R})$$

$$(\Lambda, \omega) \mapsto \Sigma, \text{ s.t. } \Sigma = \Lambda^T \Sigma \Lambda + \omega I_n.$$

ϕ_G is well defined because when $(I_{n^2} - \Lambda^T \otimes \Lambda^T)^{-1}$ is invertible,

$$\Sigma = \Lambda^T \Sigma \Lambda + \omega I_n \Leftrightarrow \text{vec}(\Sigma) = (I_{n^2} - \Lambda^T \otimes \Lambda^T)^{-1} \text{vec}(\omega I_n).$$

$$\phi_G : (\Lambda, \omega) \mapsto \Sigma, \text{ s.t. } \Sigma = \Lambda^T \Sigma \Lambda + \omega I_n.$$

The stationary VAR(1) model

By abuse of notation, define the **stationary VAR(1) model** $\mathbf{M}_G$ corresponding to a graph G as the following **set of matrices**:

$$\mathbf{M}_G = \{ \Sigma \mid \Sigma = \Lambda^T \Sigma \Lambda + \omega I_n, \Lambda \in M_G, \omega \in \mathbb{R}^+ \}.$$

$\mathbf{M}_G$ is

- the image of the parametrization map ϕ_G ,
- the set of all possible covariance matrices of the stationary distribution corresponds to a graph G .

Global identifiability

Given a **finite family** of VAR(1) models $\{\mathbf{M}_i\}_{i=1}^K$ and associated graphs $\{G_i = (V, \mathfrak{E}_{G_i})\}_{i=1}^K$, where $K \in \mathbb{N}^*$, then this family of models are identifiable if for any distinct pair (i_1, i_2) of values of i , M_{i_1} and M_{i_2} are identifiable.

Global identifiability

Two stationary VAR(1) models $\mathbf{M}_1$ and $\mathbf{M}_2$ are globally identifiable if

$$\mathbf{M}_1 \cap \mathbf{M}_2 = \emptyset.$$

- Global identifiability is too strong.

Fix $\omega = 1$,

$$\Lambda_1 = \begin{bmatrix} 0.50 & 0.70 & 0.00 \\ 0.00 & 0.90 & 0.00 \\ 0.00 & 0.80 & 0.40 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} 0.50 & 0.67 & -0.01 \\ 0.00 & 0.94 & 0.02 \\ 0.00 & 0.00 & 0.38 \end{bmatrix}$$
$$\Rightarrow \Sigma = \begin{bmatrix} 1.33 & 0.85 & 0.00 \\ 0.85 & 22.85 & 0.60 \\ 0.00 & 0.60 & 1.19 \end{bmatrix}.$$

Generic identifiability

- $\mathbf{M}_G$ admits an *algebraic dimension*.

Generic identifiability

Two stationary VAR(1) models $\mathbf{M}_1$ and $\mathbf{M}_2$ are generically identifiable if

$$\dim(\mathbf{M}_1 \cap \mathbf{M}_2) < \max\{\dim(\mathbf{M}_1), \dim(\mathbf{M}_2)\}.$$

- Models with different dimensions are generically identifiable.
- If we can calculate the dimension of the model, we can focus on the identifiability of models with the same dimension.

Model dimension and the Jacobian matrix

Recall the parametrization map:

$$\phi_G : (\Lambda, \omega) \mapsto \Sigma, \text{ s.t. } \Sigma = \Lambda^T \Sigma \Lambda + \omega I_n.$$

Jacobian matrix

The Jacobian matrix of the model is

$$\mathbf{J}_G = \left(\frac{\partial (\phi_G)_j}{\partial \theta_i} \right), 1 \leq i \leq E_G + 1, 1 \leq j \leq \frac{n(1+n)}{2},$$

where E_G is the number of edges in G , $(\phi_G)_j$'s are the distinct entries of Σ , and θ_i 's are the entries of Λ and ω .

Proposition

$$\dim(\mathbf{M}_G) = \text{rank}(\mathbf{J}_G)$$

Proof: It's sufficient to prove that ϕ_G is a rational map.

Jacobian matroid

How to deal with models with the same dimension?

Linear matroid

Let $A \in M_{m \times n}(\mathbb{R})$ be a matrix, then *the matroid defined by the column independence* of A is the pair $\{E, \mathcal{I}\}$, where $E = \{1, \dots, n\}$, representing the n columns of A , and

$$\mathcal{I} = \{S \subseteq E \mid A^S \text{ are linearly independent}\},$$

where A^S represents the set of columns of A corresponding to the coordinates S .

Jacobian matroid

The *Jacobian matroid* of the stationary VAR(1) model $\mathbf{M}_G$, denoted as $\mathcal{J}(\mathbf{M}_G)$, is the matroid defined by the column independence of the Jacobian matrix.

Characterization with Jacobian matroids

Proposition¹

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two stationary VAR(1) models. Assume that $\dim(\mathbf{M}_1) \geq \dim(\mathbf{M}_2)$. If there exists a subset S of the coordinates such that

$$S \in \mathcal{J}(\mathbf{M}_2) \setminus \mathcal{J}(\mathbf{M}_1),$$

then $\dim(\mathbf{M}_1 \cap \mathbf{M}_2) < \min\{\dim(\mathbf{M}_1), \dim(\mathbf{M}_2)\}$.

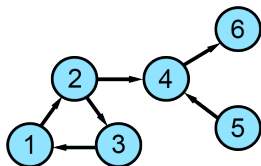
A direct result: if $\dim(\mathbf{M}_1) = \dim(\mathbf{M}_2)$, then the condition can be transformed to

$$S \in \mathcal{J}(\mathbf{M}_1) \setminus \mathcal{J}(\mathbf{M}_2) \text{ or } S \in \mathcal{J}(\mathbf{M}_2) \setminus \mathcal{J}(\mathbf{M}_1).$$

1. Sullivant, Seth. Algebraic statistics. Vol. 194. American Mathematical Society, 2023.

Identifiability Results

Maximal class - definition with illustration



The Strongly Connected Components (SCCs) of the graph are:

$$\{1, 2, 3\}, \{4\}, \{5\}, \{6\},$$

The SCCs with in-degree zero ($\{1, 2, 3\}$ and $\{5\}$) are defined as the *sources*. The set of *maximal classes* is:

$$\{\{\mathbf{1}, \mathbf{2}, \mathbf{3}, 4, 6\}, \{4, \mathbf{5}, 6\}\},$$

where the nodes in bold are sources of the respective maximal classes.

Model dimension

Theorem 1¹: model dimension

Let $\mathbf{M}_G$ be a stationary VAR(1) model where G **does not contain multi-edges**, then

$$\text{rank}(\mathbf{J}_G) = \min \{n_r, n'_c\}, \quad (1)$$

where

$$n_r = E_G + 1;$$

$$n'_c = |\{\{a, b\} \mid a, b \in [n], a, b \text{ belong to the same maximal class}\}|.$$

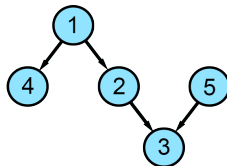
$$\mathfrak{M}\mathfrak{C} = \{\{1, 2, 3, 4\}, \{3, 5\}\};$$

$$n_r = E_G + 1 = 5 + 4 + 1 = 10;$$

$$n'_c = 5 + \binom{4}{2} + \binom{2}{2} = 12.$$

Therefore,

$$\dim(\mathbf{M}_G) = \text{rank}(\mathbf{J}_G) = \min \{n_r, n'_c\} = 10.$$



Maximal class characterization

Theorem 2¹: Models with the same dimension

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two stationary VAR(1) models s.t. $\dim(\mathbf{M}_1) = \dim(\mathbf{M}_2)$. Then $\mathbf{M}_1$ and $\mathbf{M}_2$ are generically identifiable if G_1 and G_2 have different maximal classes.

Theorem 3¹: Models with unknown dimensions

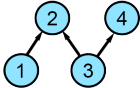
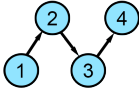
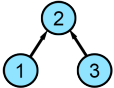
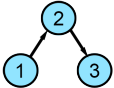
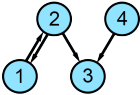
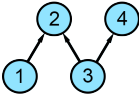
Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be two stationary VAR(1) models. Then $\mathbf{M}_1$ and $\mathbf{M}_2$ are generically identifiable if

- 1 there exist $i, j \in [n]$ s.t. i, j belong to the same maximal class in G_1 , but do not in G_2 , and
- 2 there exist $s, t \in [n]$ s.t. s, t belong to the same maximal class in G_2 , but do not in G_1 .

1. Liu, Bixuan. "Identifiability of VAR (1) model in a stationary setting." arXiv preprint arXiv:2504.03466 (2025).

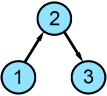
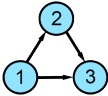
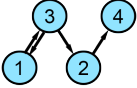
Summary of identifiability results

$\mathbf{M}_1$ and $\mathbf{M}_2$ are generically identifiable if

Condition		Example	
G_1 and G_2 do not contain multi-edges	$\dim(\mathbf{M}_1) = \dim(\mathbf{M}_2)$, and $\mathfrak{M}\mathfrak{C}_1 \neq \mathfrak{M}\mathfrak{C}_2$	G_1 	G_2 
	$\dim(\mathbf{M}_1) \neq \dim(\mathbf{M}_2)$	G_1 	G_2 
G_1 or G_2 contains multi-edges	Theorem 3 is satisfied	G_1 	G_2 

Summary of identifiability results

$\mathbf{M}_1$ and $\mathbf{M}_2$ do not satisfy the identifiability criteria in this paper if

Condition	Example
G_1 and G_2 do not contain multi-edges	$\dim(\mathbf{M}_1) = \dim(\mathbf{M}_2)$ and $\mathcal{MC}_1 = \mathcal{MC}_2$ G_1  G_2 
G_1 or G_2 contains multi-edges	Theorem 3 is not satisfied G_1  G_2 is any graph

Thank you for your attention!



Bipartite graphs with prior classification

Proposition

Any directed bipartite graphs are (generically) identifiable if the nodes are primarily classified.

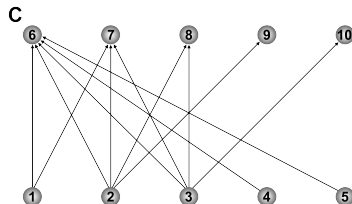


Figure: Example of ecological networks¹

The maximal classes are:

$\{\{1, 6, 7\}, \{2, 6, 7, 8, 9\}, \{3, 6, 7, 8, 9, 10\}, \{4, 8, 9\}, \{5, 10\}\}$

If we already know that species 1 – 5 are resources, and species 6 – 10 are consumers, then each maximal class contains exactly one resource and all the consumers it feeds, thus the whole graph is identifiable.

Bipartite graphs without prior classification

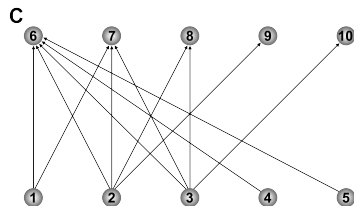


Figure: Example of ecological networks¹

The maximal classes are:

$\{\{1, 6, 7\}, \{2, 6, 7, 8, 9\}, \{3, 6, 7, 8, 9, 10\}, \{4, 8, 9\}, \{5, 9, 10\}\}$

If we do not know who are the resources, or who are the consumers, then we may not be able to recover the whole graph. But we know that each maximal class contains exactly one resource and all the consumers it feeds.

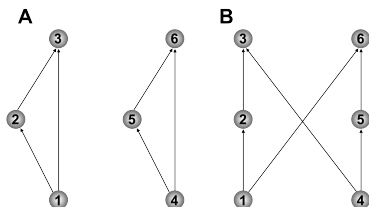
1. Ings, Montoya, Bascompte, Blüthgen, Brown, Dormann, Woodward, *Journal of animal ecology*, 2009

Resilience of the network

Maximal classes can sometimes indicate the resilience of the network.

Proposition

If two maximal classes are **disjoint**, the nodes from one maximal class are completely unrelated to the ones from the other.



Maximal classes of A:

$$\{\{1, 2, 3\}, \{4, 5, 6\}\}.$$

Maximal classes of B:

$$\{\{1, 2, 3, 6\}, \{3, 4, 5, 6\}\}.$$

Figure: Example of ecological networks¹

In this case, A is more resilient than B.

1. Ings, Montoya, Bascompte, Blüthgen, Brown, Dormann, Woodward, *Journal of animal ecology*, 2009