Identifiability of VAR(1) model in a stationary setting

Bixuan LIU

Supervised by: Stéphane Robin, Viet Chi Tran and Elisa Thébault

June 21, 2026

Overview

Introduction

Ecological background

Ecosystems are often modeled by dynamical systems.

For example, **Robert May's model**[1]

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \Lambda^T \mathbf{x}_t,$$

- $\mathbf{x}_t$: n-dimensional vector of species abundances
- Λ: deterministic $n \times n$ interaction matrix
    - cooperation, competition, predation

(and the Lotka-Volterra equations[2])

Central topic in ecology: determination of Λ.

1. May (1972) 2. Akjouj et al. (2024)

Ecological background

Central topic in ecology: determination of Λ.

Sampling insufficiency problem[1]

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 stations[2].

Subject of this work

Recovery of dynamic networks in ecosystems using only **static data**$^{(\star)}$.

$(\star)$ e.g. i.i.d. samples of abundances of the same network from multiple sites.

1. Legendre et al. (2010) 2. Chaffron et al. (2021)

VAR(1)

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

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

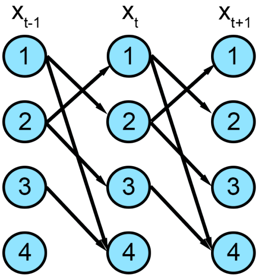
where Λ is a deterministic $n \times n$ matrix.

Corresponding graph

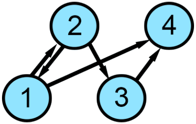
$G = (V, \mathfrak{E}_G)$, where

- $V = \{1, \cdots, n\}$, representing n interacting random variables,
- $\mathfrak{E}_G$ contains ordered pairs (i, j) s.t. $\Lambda_{ij} \neq 0$.

Graphical model



(a) The graphical VAR(1) model



(b) The encoded graph G

VAR(1) in a stationary setting

VAR(1) model:

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

If $\rho(\Lambda) < 1$, then x_t admits a stationary distribution $\mathbf{x}_\infty \sim \mathcal{N}(0, \Sigma)$ where

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

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

Obstacles

1. **Lack of reference:** VAR(1) graph recovery in the stationary regime is rarely studied; most work assumes time-series data[1,2].

2. **Classical methods not applicable:** The VAR(1) interaction graph differs from the graphical model of its stationary distribution. $\Rightarrow$ Results from Gaussian Graphical Models are irrelevant.

1. Gong et al. (2015) 2. Tank et al. (2019)

Related work

Recall

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

Young's condition[1] - 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;
- No guarantee for identifiability.

Drton's work[2]

Algebraic approaches are applied to linear structural equation models (SEMs), and sufficient graphical conditions for identifiability are derived.

1. Young et al. (2019) 2. Drton (2025)

Identifiability framework

Statistical model

The parametrization map:

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

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

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

The stationary VAR(1) model

Given a directed graph G, and

$$M_G = \{\Lambda \mid \rho(\Lambda) < 1, \text{ and } \Lambda_{ij} = 0 \text{ if } (i,j) \notin \mathfrak{E}_G\},$$

then the stationary VAR(1) model is defined as

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

- $\mathbf{M}_G$ is (i) the image of ϕ_G and (ii) the set of covariance matrices Σ that can be associated with a matrix Λ of support G.

Global identifiability

Identifiability of a finite family $\Rightarrow$ identifiability of pairs.

Global identifiability

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

$\Rightarrow$ too strong. Example: fix $\omega = 1$,

$$\Lambda_1 = \begin{bmatrix} 0.50 & 0.70 & 0.00 \\ 0.00 & 0.90 & 0.00 \\ 0.00 & 0.80 & 0.40 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} 0.50 & 0.67 & -0.01 \\ 0.00 & 0.94 & 0.02 \\ 0.00 & 0.00 & 0.38 \end{bmatrix}$$

$$\Rightarrow \Sigma = \begin{bmatrix} 1.33 & 0.85 & 0.00 \\ 0.85 & 22.85 & 0.60 \\ 0.00 & 0.60 & 1.19 \end{bmatrix},$$

$$\Rightarrow \Sigma \in \mathbf{M}_1 \cap \mathbf{M}_2.$$

Generic identifiability

- $\mathbf{M}_G$ admits a well-defined algebraic dimension.

Generic identifiability

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

Interpretation: $\mathbf{M}_1 \cap \mathbf{M}_2$ is a Lebesgue measure zero subset of $\mathbf{M}_1$ or $\mathbf{M}_2$.

Direct conclusions:

- Different dimensions $\Rightarrow$ identifiable.
- Knowing the dimension $\Rightarrow$ focusing on identifiability of models with the same dimension.

Jacobian matrix and Model dimension

The parametrization map:

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

The Jacobian matrix of the model is

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

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

Proposition

$$\dim\left(\mathbf{M}_G\right) = \operatorname{rank}\left(\mathbf{J}_G\right) \text{ at a generic point}^{(\star)}$$

Proof: It's sufficient to prove that ϕ_G is a rational map.
$\Rightarrow$ A tool to calculate model dimension.

$(\star)$ a point in a **general position**, at which all generic properties (properties that are true for almost every point) are true.

Linear matroid

Linear matroid of a matrix A

The pair $\{E, \mathcal{I}\}$, where $E = \{1, \cdots, n\}$, representing the n columns of A, and

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

Example: Let

$$A = \begin{bmatrix} 2 & 0 & 2 & 4 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 4 & 0 & 4 & 8 \end{bmatrix},$$

then the linear matroid of A is $\{E, \mathcal{I}\}$, where

- $E = \{1, 2, 3, 4\}$;
- $\mathcal{I} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\}, \{1,2,3\}, \{2,3,4\}\}$.

Jacobian matroid and identifiability criteria

Jacobian matroid $\mathcal{J}(\mathbf{M}_G)$ = the linear matroid of the Jacobian matrix.

Proposition[1]

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 (2023)

Main results

Maximal class

Definition

A maximal class of a directed graph is the set of reachable nodes from a Strongly Connected Component (SCC) with in-degree zero.

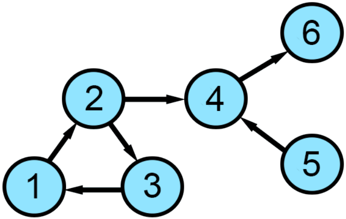
Link to algorithm: https://github.com/Bi-xuan/maximal_class
Example:

- SCCs:

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

- Sources: $\{1,2,3\}$ and $\{5\}$ (SCCs with in-degree zero);
- Set of *maximal classes* (unique):

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



Model dimension

Theorem 1: model dimension

If G *does not contain multi-edges*, then

$$\dim\left(\mathbf{M}_G\right) = \min\left\{n_r, n_c'\right\},$$

where

$$n_r = E_G + 1;$$
$$n_c' = \left|\left\{\{a,b\} \mid a, b \in [n], a, b \text{ belong to the same maximal class}\right\}\right|.$$

Maximal class characterization

Theorem 2: Models with the same dimension

$\mathbf{M}_1$ and $\mathbf{M}_2$ with the same dimension are generically identifiable if G_1 and G_2 have different maximal classes.

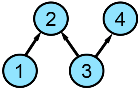
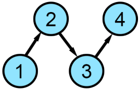
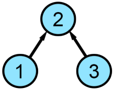
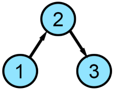
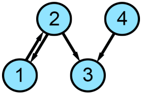
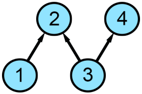
Theorem 3: Models with unknown dimensions

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

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.

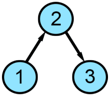
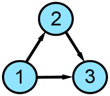
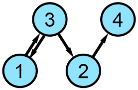
Summary

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

<table>
<tr><th colspan="2">Condition</th><th colspan="2">Example</th></tr>
<tr><td rowspan="2">G_1 and G_2 do not contain multi-edges</td><td>$\dim(\mathbf{M}_1) = \dim(\mathbf{M}_2)$, and $\mathfrak{MC}_1 \neq \mathfrak{MC}_2$</td><td>$G_1$
</td><td>G_2
</td></tr>
<tr><td>$\dim(\mathbf{M}_1) \neq \dim(\mathbf{M}_2)$</td><td>$G_1$
</td><td>G_2
</td></tr>
<tr><td>G_1 or G_2 contains multi-edges</td><td>Theorem 3 is satisfied</td><td>G_1
</td><td>G_2
</td></tr>
</table>

Summary

$\mathbf{M}_1$ and $\mathbf{M}_2$ do not satisfy the sufficient 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 $\mathfrak{MC}_1 = \mathfrak{MC}_2$	G_1 	G_2 
G_1 or G_2 contains multi-edges	Theorem 3 is not satisfied	G_1 	G_2 is any graph

Takeaway message

1. **Novelty in identifiability context:** Focus on the stationary setting rather than time series.
2. **Rigorous and extendable algebraic framework.**
3. **Practical identifiability criteria:** Connect identifiability to graph structure, yielding straightforward criteria.
4. **Relevance to ecological research:** Incorporates the sampling insufficiency problem, requiring only i.i.d. steady-state samples instead of costly time series.

Ongoing work

- **Optimization:** Determination of the support of Λ with Σ known.
- **Stats:** Estimation of the support of Λ from $\hat{\Sigma}$ (estimated from real data).

Thank you for your attention!