Identifiability of linear structural equation models with homoscedastic errors using algebraic matroids

Jun Wu[1], Mathias Drton[1], Benjamin Hollering[2]

[1]Technical University of Munich, Germany;
TUM School of Computation, Information and Technology,
Department of Mathematics and Munich Data Science Institute

[2]Research Group in Nonlinear Algebra,
Max Planck Institute for Mathematics in the Sciences

German Probabiliy and Statistics Day, March 2023

Outline

Outline

Linear Structural Equation Models

- Random vector $X = (X_i : i \in V)$ solves

$$X = \Lambda^T X + \varepsilon, \qquad \mathsf{Var}[\varepsilon] = \Omega.$$

- We consider homoscedastic errors, $\Omega = \omega \cdot I$, and then focus on the precision matrix:

$$\psi_G(\Lambda, s) = \Sigma^{-1} = s(I - \Lambda)(I - \Lambda)^T, \quad s = \frac{1}{\omega}.$$

- The linear homoscedastic Gaussian model given by a directed graph $G = (V, D)$ is

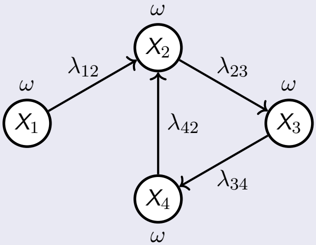
$$M_G = \left\{ s(I - \Lambda)(I - \Lambda)^T : \Lambda \in \mathbb{R}^D_{\mathrm{reg}},\ s > 0 \right\},$$

where $\mathbb{R}^D_{\mathsf{reg}} = \{\, \Lambda \in \mathbb{R}^{V \times V} : \Lambda_{ij} = 0 \ \text{ if } \ i \to j \notin D,\ I - \Lambda \text{ invertible} \}.$

Linear Structural Equation Models: Example

Example 1

$$X_1 = \varepsilon_1$$
$$X_2 = \lambda_{12}X_1 + \lambda_{42}X_4 + \varepsilon_2$$
$$X_3 = \lambda_{23}X_2 + \varepsilon_3$$
$$X_4 = \lambda_{34}X_3 + \varepsilon_4$$



$$\Lambda = \begin{pmatrix} 0 & \lambda_{12} & 0 & 0 \\ 0 & 0 & \lambda_{23} & 0 \\ 0 & 0 & 0 & \lambda_{34} \\ 0 & \lambda_{42} & 0 & 0 \end{pmatrix}, \qquad s = \frac{1}{\omega}$$

The graph is simple and the SEM is non-recursive ($\exists$ cycle)

Outline

Structural Identifiability

Question: Can two different graphs have the same model?

- Basic case: Markov equivalent classes of DAGs (directed acyclic graphs)
- Homoscedastic errors: within the class of DAGs, the graph G is known to be identifiable
  [Chen, Drton, and Wang 2019; Peters and Bühlmann 2014]
- Identifiability results when cycles allowed?

Definition

Let $\{M_i\}_{i=1}^k$ be a finite set of algebraic statistic models given by subsets of $\mathbb{R}^m$. The indices i's are generically identifiable if for each pair of (i_1, i_2),

$$\dim(M_{i_1} \cap M_{i_2}) < \max\left(\dim(M_{i_1}), \dim(M_{i_2})\right).$$

Structural Identifiability

How to compare two models?

- Traditional method: Groebner basis (equi.)
- We use: Jacobian matroid (suff., related to graphical criteria)

Our contributions

- Derive graphical criteria certifying two simple directed graphs have distinguishable models
- Give subclasses of graphs that are generically identifiable
- Computational checks for small-size graphs

For a simple directed graph $G = (V, D)$,

$$dim(M_G) := \mathrm{rank}(J(\psi_G)) = |D| + 1.$$

Outline

Jacobian: Example

Example 2

$G = (V, D)$, with $V = \{1,2,3,4\}$ and $D = \{(1,2),(2,4),(1,3),(3,4)\}$

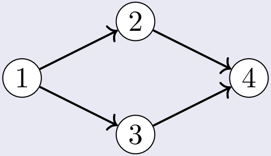


Figure: Example 2

$J(\psi_G)$:

$$
\begin{array}{cccccccccccc}
K_{11} & K_{22} & K_{33} & K_{44} & K_{12} & K_{23} & K_{34} & K_{13} & K_{24} & K_{14} & \\
2s\lambda_{12} & 0 & 0 & 0 & -s & 0 & 0 & 0 & 0 & 0 & \lambda_{12} \\
2s\lambda_{13} & 0 & 0 & 0 & 0 & 0 & 0 & -s & 0 & 0 & \lambda_{13} \\
0 & 2s\lambda_{24} & 0 & 0 & 0 & s\lambda_{34} & 0 & 0 & -s & 0 & \lambda_{24} \\
0 & 0 & 2s\lambda_{34} & 0 & 0 & s\lambda_{24} & -s & 0 & 0 & 0 & \lambda_{34} \\
1+\lambda_{12}^2+\lambda_{13}^2 & 1+\lambda_{24}^2 & 1+\lambda_{34}^2 & 1 & -\lambda_{12} & \lambda_{24}\lambda_{34} & -\lambda_{34} & -\lambda_{13} & -\lambda_{24} & 0 & s
\end{array}
$$

$$\operatorname{rank}(J_{\{44,12,34,13,24\}}) = 5$$

Jacobian Matroid

Definition

Suppose $M = \text{Im}(\phi)$ with parametrization $\phi(\theta) = (\phi_1(\theta), \ldots, \phi_r(\theta))$. Let

$$J(\phi) = \left(\frac{\partial \phi_j}{\partial \theta_i}\right), 1 \leq i \leq d, 1 \leq j \leq r$$

be the Jacobian of ϕ. Then the Jacobian matroid of model M is the matroid $\mathcal{M}(\phi) = (E, \mathcal{I})$, where

- $E = [r]$, the set of column indices
- A set $S \in \mathcal{I} \subseteq 2^E$ is called an independent set
- The columns of $J(\phi)$ indexed by S are linearly independent over the fraction field $\mathbb{R}(\theta)$

Jacobian Matroid: Example

Example 3

$$\phi(t_1, t_2, t_3) = (t_1, -t_1^2, t_1 t_2 + t_3^2),$$
$$J = \begin{bmatrix} 1 & -2t_1 & t_2 \\ 0 & 0 & t_1 \\ 0 & 0 & 2t_3 \end{bmatrix}.$$

- $E = \{1, 2, 3\}$
- The independent sets are

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1, 3\}, \{2, 3\}$$

Proving Identifiability with Algebraic Matroids

Proposition [Hollering and Sullivant 2021]

Let M_1 and M_2 be two parameterized models in $\mathbb{R}^m$ with parameterization ψ_1 and ψ_2. Assuming without loss of generality that $\dim(M_1) \geq \dim(M_2)$, if there exists a subset S of the columns such that

$$S \in \mathcal{M}(\psi_2) \setminus \mathcal{M}(\psi_1),$$

then $\dim(M_1 \cap M_2) < \min(\dim(M_1), \dim(M_2))$.

- A sufficient condition for generic identifiability
- M_1, M_2 exchangeable when $\dim(M_1) = \dim(M_2)$

Outline

Outdegree Proposition

Proposition 1 (Outdegree proposition)

Let G_1, G_2 be non-complete simple directed graphs. If $\exists i$ s.t. $|Ch_1(i)| \neq |Ch_2(i)|$ then $\mathcal{M}(\psi_1) \neq \mathcal{M}(\psi_2)$.

Example 4

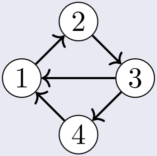


Figure: G_1

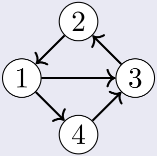


Figure: G_2

- G_1 has outdegree sequence $\{1, 1, 2, 1\}$
- G_2 has outdegree sequence $\{2, 1, 1, 1\}$

Example: Outdegree Proposition Not Applicable

Example 5

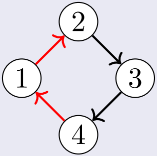


Figure: G_1

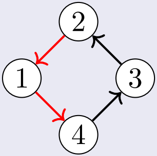


Figure: G_2

Let $S = \{22, 33, 23, 34, 14\}$,

$$J_S^1 = \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{array}{ccccc} K_{22} & K_{33} & K_{23} & K_{34} & K_{14} \end{array}$$

$$J_S^1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 2s\lambda_{23} & 0 & -s & 0 & 0 \\ 0 & 2s\lambda_{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -s \\ 1+\lambda_{23}^2 & 1+\lambda_{34}^2 & -\lambda_{23} & -\lambda_{34} & -\lambda_{41} \end{pmatrix} \begin{array}{c} \lambda_{12} \\ \lambda_{23} \\ \lambda_{34} \\ \lambda_{41} \\ s \end{array}, \quad \text{rank}(J_S^1) = 4,$$

$$\begin{array}{ccccc} K_{22} & K_{33} & K_{23} & K_{34} & K_{14} \end{array}$$

$$J_S^2 = \begin{pmatrix} 2s\lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & 2s\lambda_{32} & -s & 0 & 0 \\ 0 & 0 & 0 & -s & 0 \\ 0 & 0 & 0 & 0 & -s \\ 1+\lambda_{21}^2 & 1+\lambda_{32}^2 & -\lambda_{32} & -\lambda_{43} & -\lambda_{14} \end{pmatrix} \begin{array}{c} \lambda_{21} \\ \lambda_{32} \\ \lambda_{43} \\ \lambda_{14} \\ s \end{array}, \quad \text{rank}(J_S^2) = 5.$$

Parentally Closed set condition

Definition

A set $L \in \mathrm{ne}(i)$ is called parentally closed w.r.t node i if $\mathrm{pa}(L) \cap \mathrm{ne}(i) \subseteq L$.

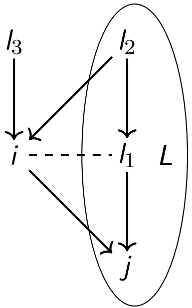


Figure: A parentally closed set

Parentally Closed Set Condition

Proposition 2 (Parentally closed set)

Let $G_1 = (V, D_1)$, $G_2 = (V, D_2)$ be two simple directed graphs, both not complete. For any node i, there are two collections of parentally closed sets $\mathcal{L}_i^1, \mathcal{L}_i^2$, corresponding to G_1 and G_2. If there is a set $L \in \mathcal{L}_i^k$ such that $|Ch_k(i) \cap L| > |Ch_{3-k}(i) \cap L|$, $k \in \{1, 2\}$, then G_1 and G_2 have different matroids.

Corollary 2.1 (Transitive triangle-free)

If two different non-complete simple graphs do not contain transitive triangles ($i \to j \to k$ and $i \to k$), then they have different matroids.

(Every parentally closed set in a transitive triangle-free graph is a singleton!)

Outline

Identifiability Results

Theorem 1

Let $\mathcal{G}$ be the collection of non-complete simple directed graphs. If the collection satisfies one of the following conditions, then the models set of graphs in $\mathcal{G}$ is generically identifiable.

(i) Every graph $G \in \mathcal{G}$ has a unique outdegree sequence

(ii) Every graph $G \in \mathcal{G}$ does not contain a transitive triangle ($i \to j \to k$ and $i \to k$)

Theorem 2

A DAG and a cyclic graph (generically) generate different distributions under homoscedastic errors condition.

Theorem 3

The collection of all non-complete simple directed graphs with at least 1 source node, and whose strongly connected components contain no transitive triangles are generically identifiable.

Outline

Computational checks for $|V| \le 6$

Methods

- $|V| \le 5$: Complete symbolic rank checks
- $|V| = 6$: Brute force check: extremely time-consuming!

To relieve the issue:

- Comparisons: within the subclasses indexed by outdegree sequences
- Parameters: random integers

Results

- Most of the simple directed graphs have unique matroids
- Some graph pairs have the same matroids, but can be distinguished by entries in the precision matrix
- Compatible with parental closed set condition checks

Conclusions

🙂

- Natural extension of equal variance (homoscedastic) error assumption from DAGs to directed cyclic graphs
- Partial identifiability results of linear homoscedastic Gaussian models
- Some side-results in algebra

🙃

- Results are not strong enough to cover all cases
- Computation capacity is currently limited to 6-node graphs

References

Chen, Wenyu, Mathias Drton, and Y Samuel Wang (Sept. 2019). "On causal discovery with an equal-variance assumption". In: *Biometrika* 106.4, pp. 973–980.

Hollering, Benjamin and Seth Sullivant (2021). "Identifiability in phylogenetics using algebraic matroids". In: *Journal of Symbolic Computation* 104, pp. 142–158.

Peters, Jonas and Peter Bühlmann (2014). "Identifiability of Gaussian structural equation models with equal error variances". In: *Biometrika* 101.1, pp. 219–228.