Identifiability of graphical models

Luis David García–Puente

Department of Mathematics and Statistics
Sam Houston State University University

AMS 2010 Spring Southeastern Sectional Meeting
Special Session on Advances in Algebraic Statistics
March 27, 2010

Joint work with Sarah Spielvogel and Seth Sullivant

CAUSALITY

- Central concept in most sciences – many physical laws describe cause–effect relationships.
- “smoking causes cancer”
- “carbon dioxide emissions contribute to global warming”
- There is not a universally agreed upon formalization of causality.
- Represent causality using **graphical models**, a representation method based on directed graphs and probability theory.

CAUSALITY

- Central concept in most sciences – many physical laws describe cause–effect relationships.
- “smoking causes cancer”
- “carbon dioxide emissions contribute to global warming”
- There is not a universally agreed upon formalization of causality.
- Represent causality using **graphical models**, a representation method based on directed graphs and probability theory.

Causality

- Central concept in most sciences – many physical laws describe cause–effect relationships.
- “smoking causes cancer”
- “carbon dioxide emissions contribute to global warming”
- There is not a universally agreed upon formalization of causality.
- Represent causality using **graphical models**, a representation method based on directed graphs and probability theory.

Causality

- Central concept in most sciences – many physical laws describe cause–effect relationships.
- “smoking causes cancer”
- “carbon dioxide emissions contribute to global warming”
- There is not a universally agreed upon formalization of causality.
- Represent causality using **graphical models**, a representation method based on directed graphs and probability theory.

Causality

- Central concept in most sciences – many physical laws describe cause–effect relationships.
- “smoking causes cancer”
- “carbon dioxide emissions contribute to global warming”
- There is not a universally agreed upon formalization of causality.
- Represent causality using **graphical models**, a representation method based on directed graphs and probability theory.

Structural Equation Models

1. The relationships among a set of observed variables are expressed by **linear equations**.

2. Each equation describes the dependence of one variable in terms of the others, and contains a **stochastic error term** accounting for the influence of unobserved factors.

3. Independence assumptions on pairs of error terms are also specified in the model.

Structural Equation Models

1. The relationships among a set of observed variables are expressed by **linear equations**.

2. Each equation describes the dependence of one variable in terms of the others, and contains a **stochastic error term** accounting for the influence of unobserved factors.

3. Independence assumptions on pairs of error terms are also specified in the model.

Structural Equation Models

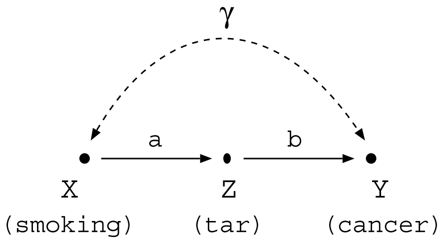
1. The relationships among a set of observed variables are expressed by **linear equations**.
2. Each equation describes the dependence of one variable in terms of the others, and contains a **stochastic error term** accounting for the influence of unobserved factors.
3. Independence assumptions on pairs of error terms are also specified in the model.

Example (Pearl 2000)

This model investigates the relations between **smoking** X and **lung cancer** Y, taking into consideration the **amount of tar** Z deposited in a person's lungs, and allowing for unobserved factors to affect both smoking X and cancer Y.

$$
\begin{aligned}
X &= \varepsilon_1 \\
Z &= aX + \varepsilon_2 \\
Y &= bZ + \varepsilon_3 \\
\mathrm{cov}(\varepsilon_1, \varepsilon_2) &= 0 \\
\mathrm{cov}(\varepsilon_2, \varepsilon_3) &= 0 \\
\mathrm{cov}(\varepsilon_1, \varepsilon_3) &= \gamma
\end{aligned}
$$

where $\varepsilon_i \sim \mathcal{N}(0, \omega_i)$.



GAUSSIAN STRUCTURAL EQUATION MODELS

Let $G = (V, D, B)$ be a graph with vertex set $V = \{1, 2, \ldots, m\}$, a set of **directed edges** D, and a set of **bidirected edges** B. Assume the subgraph of directed edges is acyclic and topologically ordered.

Let PD_n denote the set of $m \times m$ **symmetric positive definite matrices**. Let $PD(B) := \{\Omega \in PD_m : \omega_{ij} = 0 \text{ if } i \neq j \text{ and } i \leftrightarrow j \notin B\}$. Let $\epsilon \sim \mathcal{N}(0, \Omega)$ such that $\Omega \in PD(B)$.

For each $i \to j \in D$ let $\lambda_{ij} \in \mathbb{R}$ be a **parameter**. For each $j \in V$ define

$$X_j = \sum_{i : i \to j \in D} \lambda_{ij} X_i + \epsilon_j.$$

The random vector $X \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}$$

and Λ is the strictly upper triangular matrix with $\Lambda_{ij} = \lambda_{ij}$ if $i \to j \in D$ and $\Lambda_{ij} = 0$ otherwise.

Identification Problem

Decide whether the parameters in a structural model can be **determined uniquely** from the covariance matrix of the observed variables.

The identification of a model is important because, in general, no reliable quantitative conclusion can be derived from a non-identified model.

Example (Pearl 2000)

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} \omega_1 & a\omega_1 & ab\omega_1 + \gamma \\ a\omega_1 & a^2\omega_1 + \omega_2 & a^2 b\omega_1 + b\omega_2 + a\gamma \\ ab\omega_1 + \gamma & a^2 b\omega_1 + b\omega_2 + a\gamma & a^2 b^2\omega_1 + b^2\omega_2 + \omega_3 + 2ab\gamma \end{bmatrix}$$

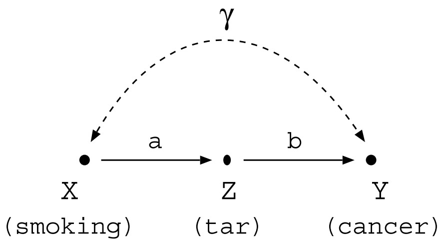
$$a = \frac{\sigma_{12}}{\sigma_{11}}$$

$$b = \frac{\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23}}{\sigma_{12}^2 - \sigma_{11}\sigma_{22}}$$

$$\omega_1 = \sigma_{11}$$

$$\omega_2 = \frac{\sigma_{11}\sigma_{22} - \sigma_{12}^2}{\sigma_{11}}$$

$$\gamma = \frac{\sigma_{11}\sigma_{12}\sigma_{23} - \sigma_{11}\sigma_{13}\sigma_{22}}{\sigma_{12}^2 - \sigma_{11}\sigma_{22}}$$



Approaches to the Identification Problem

Algebraic manipulation of the equations defining the model.

1. The method of path coefficients (Wright, 1934)
2. The rank and order criteria (Fisher, 1966)
3. Block recursive models (Fisher, 1966; Rigdon 1995)

Graphical Methods.

1. Single door criterion (Pearl, 2000)
2. Instrumental variables (Bowden and Turkington, 1984)
3. Back door criterion for total effects (Pearl, 2000)
4. G-criterion (Brito, 2006)
5. Graphical methods introduced by Tian (2004; 2005; 2007; 2009)
6. Recanting witness criterion for path-specific effects (Avin, Shpitser and Pearl, 2005)

Main Contribution

It remains unclear if these criteria (or combinations of the criteria) are necessary and sufficient to decide whether or not parameters are identifiable in a general graphical model.

Introduce a general algebraic framework for performing identifiability computations: direct effects, total effects, path-specific effects, error variances and covariances.

Identifiable Parameters

Let $\Theta \subseteq \mathbb{R}^d$ be a full dimensional **parameter set**.

Let $\mathbb{R}[\mathbf{t}] := \mathbb{R}[t_1, \ldots, t_d]$.

Let $f_1, \ldots, f_n \in \mathbb{R}[\mathbf{t}]$, and $\mathbf{f} : \Theta \to \mathbb{R}^n$ be the function defined by $\mathbf{f}(\theta) = (f_1(\theta), \ldots, f_n(\theta))^T$.

The **image** of $\mathbf{f}$ is the set $\mathbf{f}(\Theta) := \{\mathbf{f}(\theta) : \theta \in \Theta\}$.

A **parameter** is a polynomial function $u : \Theta \to \mathbb{R}$ which is not constant on Θ.

The parameter u is **identifiable** if there exists a map $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that $u(\theta) = \Phi \circ \mathbf{f}(\theta)$ for all $\theta \in \Theta$.

The parameter u is **generically identifiable** if there exists a map $\Phi : \mathbb{R}^n \to \mathbb{R}$ and a dense open subset U of Θ such that $u(\theta) = \Phi \circ \mathbf{f}(\theta)$ for all $\theta \in U$.

Given $\mathbf{f} : \Theta \to \mathbb{R}^n$ defined by $\mathbf{f}(\theta) = (f_1(\theta), \ldots, f_n(\theta))^T$, we want to check if the **parameter** u is (generically) **identifiable**.

Let $\mathbb{R}[\mathbf{p}] := \mathbb{R}[p_1, \ldots, p_n]$. The **vanishing ideal** of $S \subseteq \mathbb{R}^n$ is the set

$$\mathcal{I}(S) := \{g \in \mathbb{R}[\mathbf{p}] : g(\mathbf{a}) = 0 \text{ for all } \mathbf{a} \in S\}.$$

Let $\tilde{\mathbf{f}} = (u, f_1, \ldots, f_n)^T : \Theta \to \mathbb{R}^{d+1}$.

Let $\mathbb{R}[q, \mathbf{p}]$ be the polynomial ring with **one extra indeterminate** corresponding to the parameter function u.

Let $\mathcal{I}(\tilde{\mathbf{f}}(\Theta))$ be the vanishing ideal of the image.

Main Result

Proposition

Suppose that $g(q,\mathbf{p}) \in \mathcal{I}(\tilde{\mathbf{f}}(\Theta))$ is a polynomial such that q appears in this polynomial, $g(q,\mathbf{p}) = \sum_{i=0}^{d} g_i(\mathbf{p})q^i$ and $g_d(\mathbf{p})$ does not belong to $\mathcal{I}(\mathbf{f}(\Theta))$.

1. *If g is **linear** in q, $g = g_1(\mathbf{p})q - g_0(\mathbf{p})$ then u is **generically identifiable** by the formula $u = \frac{g_0(\mathbf{p})}{g_1(\mathbf{p})}$. If, in addition, $g_1(\mathbf{p}) \neq 0$ for $\mathbf{p} \in \mathbf{f}(\Theta)$ then u is **identifiable**.*
2. *If g has **higher degree** d in q, then u is **algebraically d-identifiable** (may or might not be identifiable).*
3. *If no such polynomial g exists then the parameter u is **not generically identifiable**.*

Computational Results

Theorem

Of the 64 graphs on three vertices,

- *there are exactly 31 graphs that are **generically identifiable** and 33 graphs that are **not generically identifiable**.*
- *The single-door criterion and instrumental variables form a **complete method** to generically identify direct causal effects for SEM models on three variables.*

Computational Results

Theorem

Of the 4096 graphs on four variables

- *exactly 1246 are **generically identifiable**, 6 are **algebraically 2-identified**, and 2844 are **not generically identifiable**.*
- *Of the 1246 generically identifiable models, exactly 1093 are generically identified by the single-door and instrumental variables criteria and the remaining 153 generically identified models contain direct causal effect parameters **only identified by the algebraic method**.*
- *There are exactly 729 bow-free models, each generically identified by the single-door criterion.*

Structural Equation Models Web Site

http://graphicalmodels.info