Learning Multivariate Causal Models

Congyuan Duan

February 14, 2022

Contents

Structure Identifiability

Methods for Structure Identification

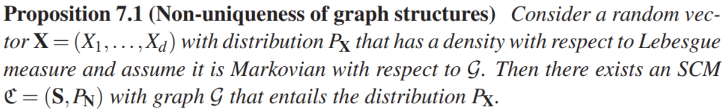
Contents

Structure Identifiability

Methods for Structure Identification

Non-uniqueness of graph structures

Given a distribution P_X over random variables $X = (X_1, \cdots, X_d)$, there is an SCM that induces the distribution P_X.

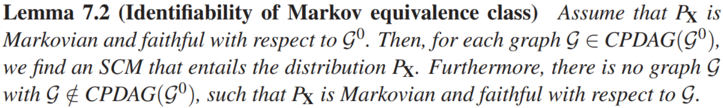


Proposition 7.1 (Non-uniqueness of graph structures) *Consider a random vector* $\mathbf{X} = (X_1, \ldots, X_d)$ *with distribution* $P_{\mathbf{X}}$ *that has a density with respect to Lebesgue measure and assume it is Markovian with respect to* $\mathcal{G}$*. Then there exists an SCM* $\mathfrak{C} = (\mathbf{S}, P_{\mathbf{N}})$ *with graph* $\mathcal{G}$ *that entails the distribution* $P_{\mathbf{X}}$*.*

In particular, given any complete DAG, we can find a corresponding SCM that entails the distribution at hand.

Identifiability of Markov equivalence class

Under the Markov condition and faithfulness, the Markov equivalence class of $\mathcal{G}_0$, represented by $CPDAG(\mathcal{G}_0)$, is identifiable from P_X.

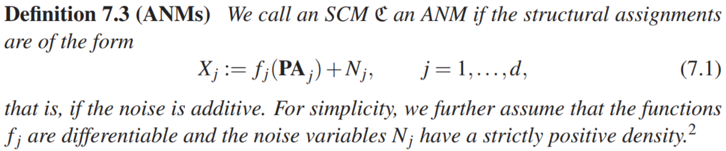


Lemma 7.2 (Identifiability of Markov equivalence class) *Assume that $P_{\mathbf{X}}$ is Markovian and faithful with respect to $\mathcal{G}^0$. Then, for each graph $\mathcal{G} \in CPDAG(\mathcal{G}^0)$, we find an SCM that entails the distribution $P_{\mathbf{X}}$. Furthermore, there is no graph $\mathcal{G}$ with $\mathcal{G} \notin CPDAG(\mathcal{G}^0)$, such that $P_{\mathbf{X}}$ is Markovian and faithful with respect to $\mathcal{G}$.*

However, we are not able to distinguish between two Markov equivalent graphs.

Additive Noise Models

We can restrict the function class to obtain non-trivial identifiability results.

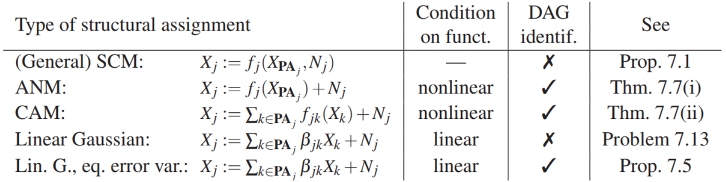


Definition 7.3 (ANMs) *We call an SCM $\mathfrak{C}$ an ANM if the structural assignments are of the form*

$$X_j := f_j(\mathbf{PA}_j) + N_j, \qquad j = 1, \ldots, d, \tag{7.1}$$

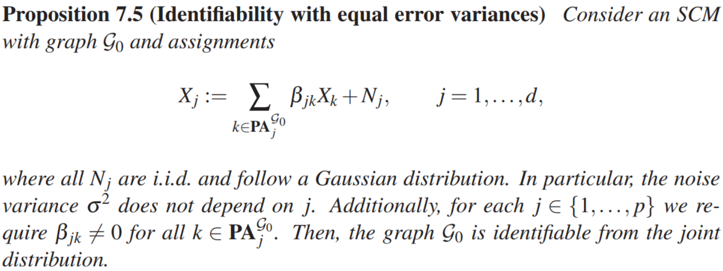
that is, if the noise is additive. For simplicity, we further assume that the functions f_j are differentiable and the noise variables N_j have a strictly positive density.[2]

Not all restricted class of SCMs described above can obtain full structure identifiability.



Type of structural assignment		Condition on funct.	DAG identif.	See
(General) SCM:	$X_j := f_j(X_{\mathbf{PA}_j}, N_j)$	—	✗	Prop. 7.1
ANM:	$X_j := f_j(X_{\mathbf{PA}_j}) + N_j$	nonlinear	✓	Thm. 7.7(i)
CAM:	$X_j := \sum_{k \in \mathbf{PA}_j} f_{jk}(X_k) + N_j$	nonlinear	✓	Thm. 7.7(ii)
Linear Gaussian:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✗	Problem 7.13
Lin. G., eq. error var.:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✓	Prop. 7.5

Linear Gaussian Models with Equal Error Variances

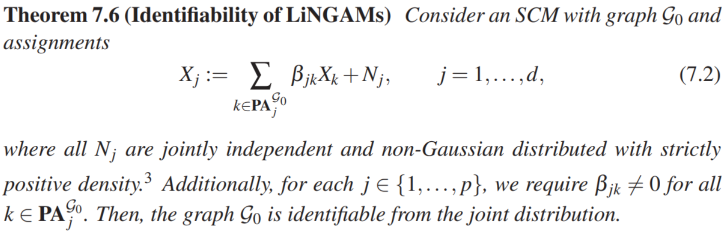


Proposition 7.5 (Identifiability with equal error variances) *Consider an SCM with graph $\mathcal{G}_0$ and assignments*

$$X_j := \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j, \qquad j = 1, \ldots, d,$$

where all N_j are i.i.d. and follow a Gaussian distribution. In particular, the noise variance σ^2 does not depend on j. Additionally, for each $j \in \{1, \ldots, p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. Then, the graph $\mathcal{G}_0$ is identifiable from the joint distribution.

Linear Non-Gaussian Acyclic Models

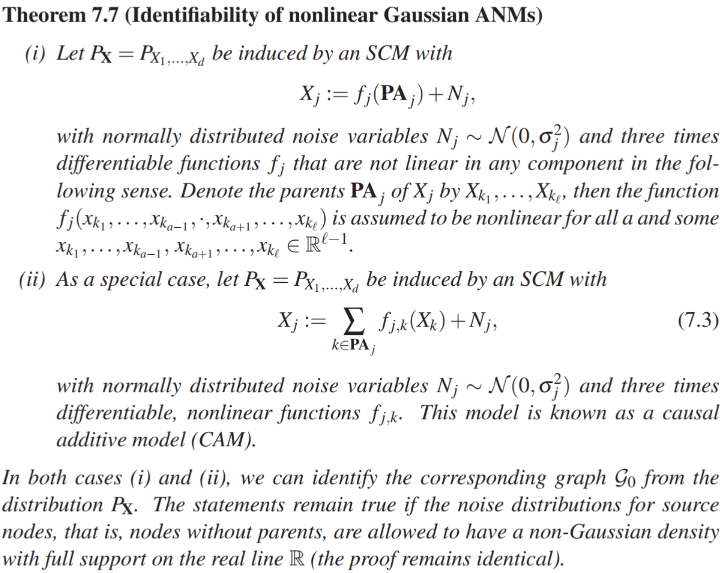


Theorem 7.6 (Identifiability of LiNGAMs) *Consider an SCM with graph $\mathcal{G}_0$ and assignments*

$$X_j := \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j, \qquad j = 1, \dots, d, \tag{7.2}$$

where all N_j are jointly independent and non-Gaussian distributed with strictly positive density.[3] *Additionally, for each $j \in \{1, \dots, p\}$, we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. Then, the graph $\mathcal{G}_0$ is identifiable from the joint distribution.*

Nonlinear Gaussian Additive Noise Models



Theorem 7.7 (Identifiability of nonlinear Gaussian ANMs)

(i) Let $P_{\mathbf{X}} = P_{X_1,\dots,X_d}$ be induced by an SCM with

$$X_j := f_j(\mathbf{PA}_j) + N_j,$$

with normally distributed noise variables $N_j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable functions f_j that are not linear in any component in the following sense. Denote the parents $\mathbf{PA}_j$ of X_j by $X_{k_1}, \dots, X_{k_\ell}$, then the function $f_j(x_{k_1}, \dots, x_{k_{a-1}}, \cdot, x_{k_{a+1}}, \dots, x_{k_\ell})$ is assumed to be nonlinear for all a and some $x_{k_1}, \dots, x_{k_{a-1}}, x_{k_{a+1}}, \dots, x_{k_\ell} \in \mathbb{R}^{\ell-1}$.

(ii) As a special case, let $P_{\mathbf{X}} = P_{X_1,\dots,X_d}$ be induced by an SCM with

$$X_j := \sum_{k \in \mathbf{PA}_j} f_{j,k}(X_k) + N_j, \tag{7.3}$$

with normally distributed noise variables $N_j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable, nonlinear functions $f_{j,k}$. This model is known as a causal additive model (CAM).

In both cases (i) and (ii), we can identify the corresponding graph $\mathcal{G}_0$ from the distribution $P_{\mathbf{X}}$. The statements remain true if the noise distributions for source nodes, that is, nodes without parents, are allowed to have a non-Gaussian density with full support on the real line $\mathbb{R}$ (the proof remains identical).

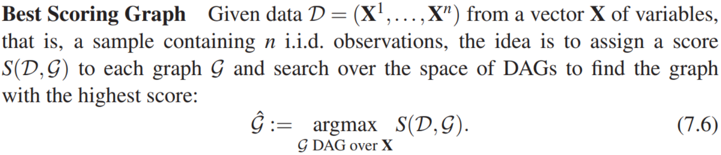
Contents

Structure Identifiability

Methods for Structure Identification

Overview

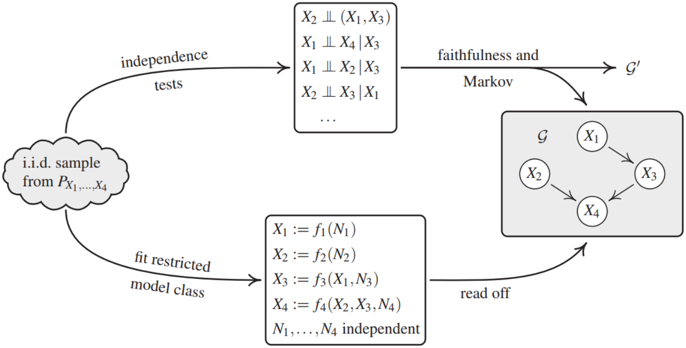
- **Independence-Based Methods:** inductive causation (IC) algorithm, PC algorithm
  Independence-based methods assume that the distribution is faithful to the underlying DAG. There is a one-to-one correspondence between d-separations in the graph and conditional independences in P_X.
- **Score-Based Methods:** additive noise models



Best Scoring Graph Given data $\mathcal{D} = (\mathbf{X}^1, \ldots, \mathbf{X}^n)$ from a vector $\mathbf{X}$ of variables, that is, a sample containing n i.i.d. observations, the idea is to assign a score $S(\mathcal{D}, \mathcal{G})$ to each graph $\mathcal{G}$ and search over the space of DAGs to find the graph with the highest score:

$$\hat{\mathcal{G}} := \operatorname*{argmax}_{\mathcal{G} \text{ DAG over } \mathbf{X}} S(\mathcal{D}, \mathcal{G}). \tag{7.6}$$

Overview



i.i.d. sample from $P_{X_1,\dots,X_4}$

independence tests

$X_2 \perp\!\!\!\perp (X_1, X_3)$

$X_1 \perp\!\!\!\perp X_4 \mid X_3$

$X_1 \perp\!\!\!\perp X_2 \mid X_3$

$X_2 \perp\!\!\!\perp X_3 \mid X_1$

$\dots$

faithfulness and Markov

$\mathcal{G}'$

$\mathcal{G}$

X_1 X_2 X_3 X_4

fit restricted model class

$X_1 := f_1(N_1)$

$X_2 := f_2(N_2)$

$X_3 := f_3(X_1, N_3)$

$X_4 := f_4(X_2, X_3, N_4)$

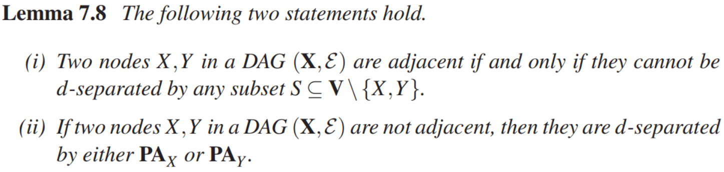
$N_1, \dots, N_4$ independent

read off

Figure 7.1: The figure summarizes two approaches for the identification of causal structures. Independence-based methods (top) test for conditional independences in the data; these properties are related to the graph structure by the Markov condition and faithfulness. Often, the graph is not uniquely identifiable; the method may therefore output different graphs $\mathcal{G}$ and $\mathcal{G}'$. Alternatively, one may restrict the model class and fit the SCM directly (bottom).

Independence-Based Methods

Most independence-based methods first estimate the skeleton, that is, the undirected edges, and orient as many edges as possible afterward.



Lemma 7.8 *The following two statements hold.*

(i) Two nodes X,Y in a DAG $(\mathbf{X},\mathcal{E})$ are adjacent if and only if they cannot be d-separated by any subset $S \subseteq \mathbf{V} \setminus \{X,Y\}$.

(ii) If two nodes X,Y in a DAG $(\mathbf{X},\mathcal{E})$ are not adjacent, then they are d-separated by either $\mathbf{PA}_X$ or $\mathbf{PA}_Y$.

We should be able to orient the v-structures in the graph. Suppose that the skeleton contains the structure $X - Z - Y$ with no direct edge between X and Y; further, let A be a set that d-separates X and Y. The structure is a v-structure if and only if $Z \notin A$.

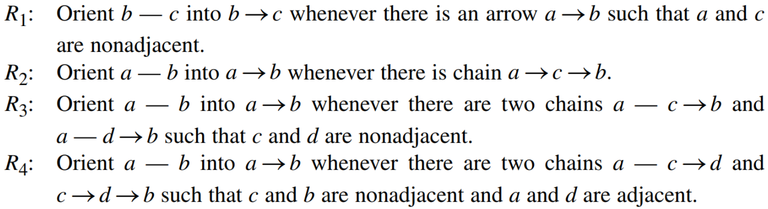
IC algorithm

Abstractly, the algorithm works as follows:

- Start with a complete undirected graph on all variables.
- For each pair of variables, see if conditioning on some set of variables makes them conditionally independent; if so, remove their edge.
- Identify all colliders by checking for conditional dependence; orient the edges of colliders.
- Try to orient undirected edges by consistency with already-oriented edges; do this recursively until no more edges can be oriented.

IC algorithm

In the last step, the following four rules are required for obtaining a maximally oriented pattern.



R_1: Orient b — c into $b \to c$ whenever there is an arrow $a \to b$ such that a and c are nonadjacent.

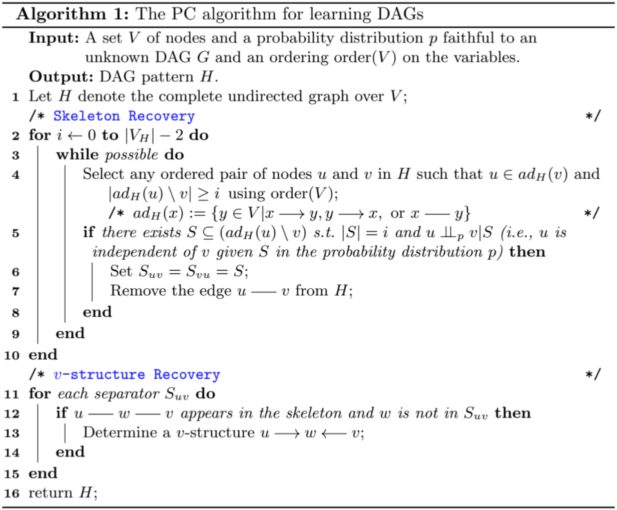
R_2: Orient a — b into $a \to b$ whenever there is chain $a \to c \to b$.

R_3: Orient a — b into $a \to b$ whenever there are two chains a — $c \to b$ and a — $d \to b$ such that c and d are nonadjacent.

R_4: Orient a — b into $a \to b$ whenever there are two chains a — $c \to d$ and $c \to d \to b$ such that c and b are nonadjacent and a and d are adjacent.

PC algorithm

Searching through all possible subsets A does not seem optimal, especially if the graph is sparse. The PC algorithm step-by-step increases the size of the conditioning set A.

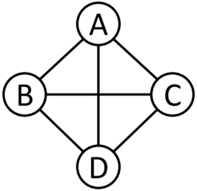


Algorithm 1: The PC algorithm for learning DAGs

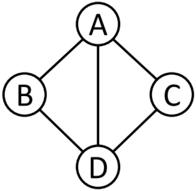
```
Input: A set $V$ of nodes and a probability distribution $p$ faithful to an
       unknown DAG $G$ and an ordering order($V$) on the variables.
Output: DAG pattern $H$.
Let $H$ denote the complete undirected graph over $V$;
/* Skeleton Recovery                                                     */
for $i \leftarrow 0$ to $|V_H| - 2$ do
    while possible do
        Select any ordered pair of nodes $u$ and $v$ in $H$ such that $u \in ad_H(v)$ and
            $|ad_H(u) \setminus v| \geq i$ using order($V$);
            /* $ad_H(x) := \{y \in V | x \longrightarrow y, y \longrightarrow x$, or $x \longrightarrow y\}$   */
        if there exists $S \subseteq (ad_H(u) \setminus v)$ s.t. $|S| = i$ and $u \perp\!\!\!\perp_p v | S$ (i.e., $u$ is
         independent of $v$ given $S$ in the probability distribution $p$) then
            Set $S_{uv} = S_{vu} = S$;
            Remove the edge $u$ —— $v$ from $H$;
        end
    end
end
/* v-structure Recovery                                                  */
for each separator $S_{uv}$ do
    if $u$ —— $w$ —— $v$ appears in the skeleton and $w$ is not in $S_{uv}$ then
        Determine a $v$-structure $u \longrightarrow w \longleftarrow v$;
    end
end
return $H$;
```

PC algorithm: example

Obtained conditional independence tests: $B \perp\!\!\!\perp C | A$, $A \perp\!\!\!\perp D | (B, C)$.
Start with the fully connected undirected graph

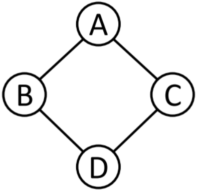


After $i = 1$

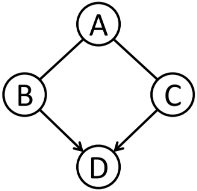


PC algorithm: example

After $i = 2$



v-structure recovery



Score-Based Methods

- **Score Function:** In the nonlinear Gaussian case, for a given graph structure $\mathcal{G}$, we regress each variable on its parents and obtain the score

$$S(\mathcal{D}, \mathcal{G}) := \log p(\mathcal{D}|\hat{\theta}, \mathcal{G}) - \frac{\#parameters}{2} \log n$$

$$\log p(\mathcal{D}|\mathcal{G}) = \sum_{j}^{d} - \log \widehat{var}[R_j]$$

  here, $\widehat{var}[R_j]$ is the empirical variance of the residuals R_j obtained from the regression of variable X_j on its parents.

- **Greedy Search Techniques:** At each step there is a candidate graph and a set of neighboring graphs. For all these neighbors, one computes the score and considers the best-scoring graph as the new candidate. If none of the neighbors obtains a better score, the search procedure terminates.

Reference

Pearl J. Causality[M]. Cambridge university press, 2009.
https://www.stat.cmu.edu/ cshalizi/402/lectures
https://pooyanjamshidi.github.io/csce580/lectures/