Additive Noise Models: Identifiability, Learning Algorithms, Hidden Variables and Time Series

Behrad Moniri

Sharif University of Technology

bemoniri@ee.sharif.edu

January 22, 2019

Model Definition

We call the SCM, C, an addtive noise model if each observed variable X_j is associated with a node j in a directed acyclic graph G, and the value of X_j is obtained as a function of its parents in G, plus independent additive noise N_j, i.e.

$$X_j = f_j(\mathbf{PA}_j) + N_j\,, \qquad j = 1, \dots, p \tag{1}$$

with jointly independent variables N_j. We will assume that the noise variables have strictly positive density.

Causal Minimality

For those models with strictly positive density, causal minimality reduces to the condition that each function f_j is not constant in any of its arguments.

Identifiability: The Bivariate Case

Hoyer et al. (2009) proves the following theorem about the identifiability of bivariate additive noise models.

Identifiability: The Bivariate Case

Theorem

An additive noise model with two variables, i.e., $X_1 = N_1$ and $X_2 = f_j(X_1) + N_2$, with $N_1 \perp\!\!\!\perp N_2$, is identifiable if it does not solve the following differential equation for all x_i, x_j with $\nu''(x_j - f(x_i))f'(x_i) \neq 0$:

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'},$$

Here $\xi := \log p_{X_1}$ and $\nu := \log p_{N_2}$ and we have skipped the arguments $x_2 - f(x_1)$, x_1, and x_1 for ν, ξ, and f and their derivatives, respectively.

Corollary

Gaussian Noise: *Assume that* $\nu''' = \xi''' = 0$ *everywhere. If a backward model exists, then* f *is linear.*

Corollary

Assume that $f(x) = x$ *and* $p_x(x) = e^{-x-e^{-x}}$ *and* $p_n(n) = e^{-n-e^{-n}}$. *With* $p_y(y) = e^{-y-2\log(1+e^y)}$, $\tilde{p}_n(\tilde{n}) = e^{-2\tilde{n}-e^{-\tilde{n}}}$ *and* $g(y) = \log(1+e^{-y})$, *one obtains:*

$$p(x,y) = p_n(y - f(x))p_x(x) = \tilde{p}_n(x - g(y))p_y(y)$$

so the model is not identifiable.

Identifiability: From Bivariate to Multivariate

Definition

Consider an ANM with p variables. We call this SCM a *restricted additive noise model* if for all $j \in \mathbf{V}$, $i \in \mathbf{PA}_j$ and all sets $\mathbf{S} \subseteq \mathbf{V}$ with $\mathbf{PA}_j \setminus \{i\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_j \setminus \{i,j\}$, there is an $x_\mathbf{S}$ with $p_\mathbf{S}(x_\mathbf{S}) > 0$, such that

$$\left(f_j(x_{\mathbf{PA}_j \setminus \{i\}}, X_i), \mathcal{L}(X_i \mid X_\mathbf{S} = x_\mathbf{S}), \mathcal{L}(N_j) \right)$$

satisfies the bivariate identifiability conditions.
We assume that the noise variables to have non-vanishing densities and the functions f_j are three times differentiable.

Identifiability: The Multivariate Case

Peters et al. (2014) proves a very interesting theorem. This theorem states how we can generalize a bivariate identifiability to the multivariate case, in this case ANM identifiability.

Theorem

Let $X_1, \ldots, X_p$ be generated by a restricted additive noise model with graph G_0 and assume that $P_{\mathbf{X}}$ satisfies causal minimality with respect to G_0, i.e., the functions f_j are not constant. Then, G_0 is identifiable from the joint distribution.

Identifiability: Post Non Linear (PNL) Models

Definition

PNL Models are introduced in Zhang and Hyvärinen (2009). A PNL is an SCM where each expresses each variable X_i as

$$X_i = g_i(f_i(\mathbf{PA}_i) + N_i), \quad i = 1, \ldots, n$$

Theorem (Bivariate Identifiability)

Assume that $x_2 = f_2(f_1(x_1) + e2)$ and $x_1 = g_2(g_1(x_2) + e1)$. *Densities and nonlinear functions are three times differentiable. We then have the following equation for every* (x_1, x_2) *satisfying* $\eta'' h' \neq 0$*:*

$$t_1 = g_2^{-1}(x_1),\ z_2 = f_2^{-1}(x_2),\ h = f_1 \circ g_2,\ h_1 = g_1 \circ f_2$$

$$\eta_1(t_1) = \log p_{t_1}(t_1)\ \ \eta_2(e_2) = \log p_{e_2}(e_2)$$

$$\eta_1''' - \frac{\eta_1'' h''}{h'} = \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\right).h' h'' - \frac{\eta_2'''}{\eta_2''}.h' \eta_1'' + \eta_2'.\left(h''' - \frac{h''^2}{h'}\right)$$

and h_1 *depends on* η_1*,* η_2*, and* h *in the following way:*

$$\frac{1}{h_1'} = \frac{\eta_1'' + \eta_2'' h'^2 - \eta_2' h''}{\eta_2'' h'}$$

	p_{e_2}	p_{t_1} $(t_1 = g_2^{-1}(x_1))$	$h = f_1 \circ g_2$	Remark
I	Gaussian	Gaussian	linear	h_1 also linear
II	log-mix-lin-exp	log-mix-lin-exp	linear	h_1 strictly monotonic, and $h_1' \to 0$, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp	one-sided asymptotically exponential (but not log-mix-lin-exp)	h strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$	—
IV	log-mix-lin-exp	generalized mixture of two exponentials	Same as above	—
V	generalized mixture of two exponentials	two-sided asymptotically exponential	Same as above	—

Figure: All unidentifiable cases with the assumptions made above

Score Based Method

RESIT Algorithm

- First proposed in Peters et al. (2014)
- Assumption : Multivariate ANM + Causal Sufficiency
- Idea : $X_i \text{ is sink } \iff N_i \perp\!\!\!\perp \mathbf{X} \setminus \{X_i\}$
- The are two stages in the algorithm:
  - Stage 1 : Finding a causal order
  - Stage 2 : Estimating DAG by removing edges
- Number of Tests (Less than PC)
  - Stage 1 : $O(n^2)$
  - Stage 2 : $O(n)$

RESIT Algorithm

Algorithm 1 Regression with subsequent independence test (RESIT)

Input: I.i.d. samples of a p-dimensional distribution on $(X_1, \dots, X_p)$
$S := \{1, \dots, p\}, \pi := [\ ]$
PHASE 1: Determine topological order.
repeat
  for $k \in S$ **do**
    Regress X_k on $\{X_i\}_{i \in S \setminus \{k\}}$.
    Measure dependence between residuals and $\{X_i\}_{i \in S \setminus \{k\}}$.
  end for
  Let k^* be the k with the weakest dependence.
  $S := S \setminus \{k^*\}$
  $\mathrm{pa}(k^*) := S$
  $\pi := [k^*, \pi]$ (π will be the topological order, its last component being a sink)
until $\#S = 0$
PHASE 2: Remove superfluous edges.
for $k \in \{2, \dots, p\}$ **do**
  for $\ell \in \mathrm{pa}(\pi(k))$ **do**
    Regress $X_{\pi(k)}$ on $\{X_i\}_{i \in \mathrm{pa}(\pi(k)) \setminus \{\ell\}}$.
    if residuals are independent of $\{X_i\}_{i \in \{\pi(1), \dots, \pi(k-1)\}}$ **then**
      $\mathrm{pa}(\pi(k)) := \mathrm{pa}(\pi(k)) \setminus \{\ell\}$
    end if
  end for
end for
Output: $(\mathrm{pa}(1), \dots, \mathrm{pa}(p))$

RESIT Algorithm : Performance (Linear Setting)

$$\beta_{jk} \sim [-2,-0.1] \cup [0.1,2] \qquad N_j \sim K_j \cdot \mathrm{sign}(M_j) \cdot |M_j|^{\alpha_j} \text{ such that } M_j \sim N(0,1),$$
$$K_j \sim U(0.1, 0.5) \text{ and } \alpha_j \sim U([2,4]).$$

	GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
	$p=4, n=100$							
DAG	0.7 ± 0.9	0.6 ± 0.8	1.2 ± 1.3	1.9 ± 1.2	3.5 ± 1.5	3.6 ± 1.4	3.1 ± 1.7	4.4 ± 1.0
CPDAG	1.1 ± 1.5	0.9 ± 1.4	1.5 ± 1.7	2.4 ± 1.5	2.4 ± 1.7	2.3 ± 1.6	2.0 ± 2.0	4.3 ± 1.4
	$p=4, n=500$							
DAG	0.2 ± 0.6	0.1 ± 0.3	0.6 ± 0.8	0.5 ± 0.8	3.1 ± 1.4	3.2 ± 1.4	2.9 ± 1.6	4.1 ± 1.2
CPDAG	0.3 ± 0.9	0.2 ± 0.5	0.9 ± 1.3	0.8 ± 1.2	1.9 ± 1.8	1.6 ± 1.7	1.6 ± 1.9	3.9 ± 1.4
	$p=15, n=100$							
DAG	12.2 ± 5.3	–	25.2 ± 8.3	11.1 ± 3.7	13.0 ± 3.6	13.7 ± 3.7	12.7 ± 4.2	57.4 ± 26.4
CPDAG	13.2 ± 5.4	–	27.0 ± 8.5	12.4 ± 3.9	10.7 ± 3.5	10.8 ± 3.8	12.4 ± 4.9	58.5 ± 27.1
	$p=15, n=500$							
DAG	6.1 ± 6.4	–	51.2 ± 17.8	3.4 ± 2.8	10.2 ± 3.8	10.8 ± 4.2	8.7 ± 4.6	57.6 ± 24.2
CPDAG	6.8 ± 6.9	–	54.5 ± 18.5	4.5 ± 3.8	8.2 ± 4.6	7.5 ± 4.4	7.1 ± 5.6	58.9 ± 25.0

Figure: Structural Hamming Distance of Estimated Graph

RESIT Algorithm : Performance (Non Linear Setting)

Functions sampled from a Gaussian process with $BW = 1$. Gaussian Noise with random variance.

	GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
	$p = 4, n = 100$							
DAG	1.5 ± 1.4	1.0 ± 1.0	1.7 ± 1.3	3.5 ± 1.2	3.5 ± 1.5	3.8 ± 1.4	3.5 ± 1.3	4.0 ± 1.3
CPDAG	1.7 ± 1.7	1.2 ± 1.4	2.0 ± 1.6	3.0 ± 1.4	2.9 ± 1.5	2.7 ± 1.4	3.4 ± 1.7	3.9 ± 1.4
	$p = 4, n = 500$							
DAG	0.5 ± 0.9	0.3 ± 0.5	0.8 ± 0.9	3.7 ± 1.2	3.5 ± 1.5	3.8 ± 1.5	3.3 ± 1.5	4.1 ± 1.2
CPDAG	0.6 ± 1.1	0.6 ± 1.0	1.0 ± 1.3	3.0 ± 1.7	3.1 ± 1.9	2.8 ± 1.8	3.4 ± 1.9	3.8 ± 1.6
	$p = 15, n = 100$							
DAG	14.3 ± 4.9	–	15.4 ± 5.7	15.4 ± 3.6	14.2 ± 3.5	15.5 ± 3.6	24.8 ± 6.3	56.8 ± 24.1
CPDAG	15.1 ± 5.4	–	16.5 ± 5.9	15.3 ± 4.0	13.3 ± 3.6	13.3 ± 4.0	26.4 ± 6.5	58.0 ± 24.7
	$p = 15, n = 500$							
DAG	13.0 ± 8.4	–	10.1 ± 5.7	21.4 ± 6.9	13.9 ± 4.5	15.1 ± 4.8	26.8 ± 8.5	56.1 ± 26.8
CPDAG	14.2 ± 9.2	–	11.3 ± 6.3	21.1 ± 7.3	13.7 ± 4.9	13.4 ± 5.1	28.6 ± 8.8	57.0 ± 27.3

Figure: Structural Hamming Distance of Estimated Graph

References

Patrik O. Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696. 2009.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15(1):2009–2053, 2014.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647–655, 2009.

The End