A linear nongaussian acyclic model (LiNGAM) for causal discovery

Aapo Hyvärinen
with
Shohei Shimizu, Yutaka Kano, and Patrik O. Hoyer

University of Helsinki / University of Osaka

[ Presentation at Osaka Workshop, Sep 2005 ]
The ‘causal discovery’ problem

- Example: is smoking cause, effect or both?
- A probabilistic model of the data allows you to predict one quantity from observation of the other
- A causal model would allow you to predict the effect on one variable if intervening on the other

![Graph showing physiological quantity vs smoking status](image-url)
The ‘causal discovery’ problem

- We observe a set of variables $x, y, z, \ldots$

- Causal discovery goal: Find the data generating mechanism, not just $p(x, y, z, \ldots)$

- In Pearl’s notation: we want to be able to quantify expressions such as $p(y \mid \text{do}(x))$ in addition to $p(y \mid x)$
How to “best” do it?

- Randomized experiments!
- Unfortunately: in many cases, can be...
  - costly
  - unpractical
  - unethical
  ...what then?
Assumption-based causal discovery from non-experimental data  (see, e.g. Spirtes et al, 1993; Pearl, 2000)

- Make some reasonable assumptions on the data-generating process
- If the assumptions are well-chosen, they allow us to infer the causal connections and directions
- Statistical methods cannot determine causality in general, but if the assumptions hold, we can choose between the alternatives present.
Novelty of LiNGAM
(Linear Non-Gaussian Acyclic Model)

• As usual, we assume a DAG, linearity and causal sufficiency, but...

• We propose to replace the gaussian assumption with that of non-gaussianity, which
  - is sometimes more realistic
  - leads to all parameters of the model being identified
First assumption: linear SEM

- A linear structural equation model (SEM):
  \[ x_i = \sum_{i \neq j} b_{ij} x_j + e_i \quad \text{or} \quad x = Bx + e \]

- Graphical representation where each observed variable corresponds to a node, and arrow between \( x_i \) and \( x_j \) corresponds to non-zero \( b_{ij} \)

- In machine learning, also called “graphical model” / “Bayesian network”
Examples

\[ x_4 := e_4 \]
\[ x_2 := 0.2x_4 + e_2 \]
\[ x_1 := x_4 + e_1 \]
\[ x_3 := -2x_2 - 5x_1 + e_3 \]

\[ x_2 := e_2 \]
\[ x_3 := 4x_2 + e_3 \]
\[ x_1 := -3x_2 + 2x_3 + e_1 \]
Assumption of acyclicility
(DAG, recursivity)

- We assume the graph defined by SEM is acyclic: there is no path from a variable $x_i$ to itself, i.e., a cycle.

- Path means moving from a node (variable) to another so that move from $x_i$ to $x_j$ is permitted if and only if $b_{ij}$ is not zero.

- Called *recursivity* in SEM, *DAG* (directed acyclic graph) in machine learning.
Note: Acyclicity and ordering

- Acyclicity is equivalent to existence of an ordering of the variables so that there are only arrows “forward”. This we call a **causal order**.

- Simple to find such ordering:
  Set $n=1$. Repeat until no nodes left:
  - Find a node with no incoming arrows. Call it number $n$.
  - Delete it and arrows from it. Increase $n$ by 1.

- When we re-order the variables according to such ordering, the matrix $B$ becomes lower triangular.
LiNGAM assumptions total:

1. The observed variables can be arranged into a causal order $k(i)$: no later variable causes any earlier variable (i.e. DAG / recursive)

2. Each variable is linear combination of “preceding” variables, plus a disturbance:

   $$ x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i $$

3. Disturbances are mutually independent (‘causally sufficient’)

4. Disturbances are non-gaussian (non-normally distributed)
Fundamental note:

- If the disturbance variables were gaussian, this would just be regular SEM.
- Then it would be impossible to estimate all parameters in general: Several models are equivalent!
- But: **non-gaussianity** allows us to estimate the full structure!
Basic insight

• All observed variables are linear combinations of the disturbance variables, i.e. we have

\[ x = Bx + e \quad \text{and} \quad x = (I - B)^{-1}e = Ae \]

• The disturbance variables \( e_i \) are non-gaussian and independent

...hence, we have a classic case of

Independent Component Analysis (ICA)
Independent Component Analysis
(Jutten and Hérault, 1991; Hyvärinen et al, 2001)

- Nongaussian version of factor analysis (or factor rotation):

\[ x_i = \sum_j a_{ij} s_j \text{ or } x = As \]

- The factors / components \( s_j \) are assumed to be nonnormal and mutually independent

- No noise, and same dimensions for \( s_j \) and \( x_i \)

- The coefficients / loadings \( a_{ij} \) can be estimated without any additional assumptions! (Comon, 1994)
ICA separates signals

Original inspiration for ICA: can we recover signals from linear mixtures?
ICA separates signals
For more information on independent component analysis:
Permutation problem in ICA

• The order of the factors is not defined. ICA gives $W = A^{-1} = I - B$ for some random permutation of its rows.

• Is there a “right” permutation, how to find it?

• Lemma: because the model is recursive, only one permutation gives $W$ with all non-zero diagonal entries (if no estimation errors).

• In practice, find permutation that gives largest values in diagonal. Can be justified as MLE. Computationally feasible (linear programming).
Finding causal order

- After obtaining estimate of $B$, we find the causal order by finding a permutation $P$ so that $PBP^T$ is as close to lower triangular as possible.

- We define an objective function as simply the sum of squares above the diagonal.

- Minimization is a very difficult problem, no efficient methods known.

- Afterwards, we can assume the entries above diagonal are just errors and fix them to zero.
Complete LiNGAM method

- Using a sample of observed data vectors $\mathbf{x}$, estimate $\mathbf{W} = \mathbf{A}^{-1}$ using standard ICA.

- Find the appropriate permutation of rows of $\mathbf{W}$, such that the absolute sum of diagonal elements is maximized.

- Divide each row by its diagonal element, then calculate $\mathbf{B} = \mathbf{I} - \mathbf{W}$.

- To find causal order: Find a permutation (same for rows and columns) which makes $\mathbf{B}$ as close to lower triangular as possible.

- Set upper triangular elements to zero.
Code

We distribute full Matlab/Octave code for LiNGAM. Please see:

http://www.cs.helsinki.fi/group/neuroinf/lingam/
Experiments on artificial data

![Graph showing relationships between number of variables and number of data vectors for different numbers of variables (3, 5, 7). The graphs show the scatter plots of generating $b_{ij}$ versus estimated $b_{ij}$ for different numbers of data vectors (100, 1000, 10000).]
Experiments on time series

- Testbed: Real causal direction known for \((x(t), x(t-1), \ldots x(t-k))\)

- In most cases LiNGAM finds the right direction, e.g. AUD vs USD exchange rate

- Sometimes the method finds inverse direction: Due to nonstationarity, not a linear model?
Summary

- Causal discovery from non-experimental data is possible by making general assumptions on causal structure.

- For continuous-valued data, common assumptions are linearity and normality, but this leads to several indistinguishable models.

- Linearity + non-normality allows the full model, including all parameters, to be estimated.

- Basic method: ICA + permutations.

- We provide full Matlab/Octave code package.