Linear Non-Gaussian Acyclic Model for Causal Discovery

Aapo Hyvärinen

Dept of Computer Science
University of Helsinki, Finland

_with_

Patrik Hoyer, Shohei Shimizu, Kun Zhang, Steve M. Smith
Estimating causal direction is fundamental problem in science
Bayesian networks or structural equation models (SEM) are ill-defined for gaussian data
For non-Gaussian data, SEM is identifiable (Shimizu et al, JMLR 2006)
Theory closely related to independent component analysis (ICA)
A simple approach possible based on likelihood ratios of variable pairs (Hyvärinen and Smith, JMLR, 2013)
Practical models for causal discovery

- Model connections between the measured variables: Which variable causes which?
- “Discovery” means data-driven approach
- “Correlation does not equal causation”: but we can go beyond correlation
- Two fundamental approaches
  - If we have time series and time-resolution of measurements fast enough:
    - we may be able to use autoregressive modelling (Granger causality)
  - Otherwise, use structural equation models (here)
Structural equation models

How does an externally imposed change in one variable affect the others?

Assume influences are linear, and all variables observable:

\[ x_i = \sum_{j \neq i} b_{ij} x_j + e_i \text{ for all } i \]

Difficult to estimate, not simple regression

Classic methods fail: not identifiable

Becomes identifiable if data non-Gaussian (Shimizu et al., JMLR, 2006)
Starting point: Two variables

- Consider two random variables, $x$ and $y$, both standardized (zero mean, unit variance)
- Goal: distinguish between two statistical models:

\[ y = \rho x + d \quad (x \to y) \quad (1) \]
\[ x = \rho y + e \quad (y \to x) \quad (2) \]

where disturbances $d, e$ are independent of $x, y$.

- If variables gaussian, completely symmetric:
  - Variance explained same for both models
  - Likelihood same for both models (simple function of $\rho$)
Non-Gaussianity comes to rescue

Real-life signals often non-Gaussian
Assumption of non-Gaussianity

- We assume that in each model, regressor or residual or both are non-Gaussian

\[ y = \rho x + d \quad (x \rightarrow y) \]  
\[ x = \rho y + e \quad (y \rightarrow x) \]

where disturbances \( d, e \) are independent of \( x, y \).

- Non-Gaussianity breaks the symmetry between \( x, y \) (Dodge and Rousson, 2001; Shimizu et al, 2006).

- We can just compare the likelihoods of the models.
Illustration of symmetry-breaking

non-Gaussian

Gaussian
Intuitive idea behind non-Gaussianity

- Central limit theorem: sums of independent variables tend to be more Gaussian
- Assume (just on this slide!) that residuals are Gaussian
  - For $y = \rho x + d$, $y$ must be more gaussian than $x$
  - So, causality must be from the less Gaussian variable to the more Gaussian
- We could measure non-Gaussianity with classical measures, e.g. kurtosis/skewness and just look at the difference of kurtoses of $x$ and $y$. 
Intuitive idea behind non-Gaussianity

- Central limit theorem: sums of independent variables tend to be more Gaussian
- Assume (just on this slide!) that residuals are Gaussian
  - For $y = \rho x + d$, $y$ must be more gaussian than $x$
  - So, causality must be from the less Gaussian variable to the more Gaussian
- We could measure non-Gaussianity with classical measures, e.g. kurtosis/skewness and just look at the difference of kurtoses of $x$ and $y$.
- This is a simple illustration with its flaws
  - The method fails for non-Gaussian residuals
  - Kurtosis/skewness not a good measure of non-Gaussianity in terms of classical statistical measures (asymptotic variance, robusteness)
Likelihood ratio and non-Gaussianity

- Principled approach (Hyvärinen and Smith, JMLR, 2013)
- Ratio of probabilities that data comes from the two models
- Asymptotic limit of the log-likelihood ratio

\[
\lim \log \frac{L(x \rightarrow y)}{L(y \rightarrow x)} = -H(x) - H(d/\sigma_d) + H(y) + H(e/\sigma_e)
\]

with \(H\), differential entropy; residuals \(d = y - \rho x\), \(e = x - \rho y\) with variances \(\sigma_d^2\), \(\sigma_e^2\).

- Entropy is maximized by Gaussian distribution
- Log-likelihood ratio is thus

\[
\text{nongaussianity}(x) + \text{nongaussianity}(\text{residual } x \rightarrow y) - \text{nongaussianity}(y) - \text{nongaussianity}(\text{residual } y \rightarrow x)
\]
Likelihood ratios and independence

- We can equally interpret the likelihood ratio as independence
- We had asymptotic limit of the likelihood ratio as

\[
\frac{\log L(x \rightarrow y)}{\log L(y \rightarrow x)} = -H(x) - H(d/\sigma_d) + H(y) + H(e/\sigma_e) \tag{5}
\]

with \(H\), differential entropy;
residuals \(d = y - \rho x\), \(e = x - \rho y\) with variances \(\sigma_d^2, \sigma_e^2\).
- Mutual information \(I(u, v) = H(u) + H(v) - H(u, v)\)
measures statistical dependence
- Log-likelihood ratio can be manipulated to give

\[
I(y, e) - I(x, d) \tag{6}
\]
since the terms related to \(H(x, e)\) and \(H(y, e)\) cancel.
Even simpler approximation of likelihood ratios

- We can make first-order approximations to obtain:

\[
\frac{\log L(x \rightarrow y)}{\log L(y \rightarrow x)} \approx \frac{\rho}{T} \sum_t -x_t g(y_t) + g(x_t) y_t
\]

where typically \( g(u) = -\tanh(u) \) and \( \rho \) is the correlation coefficient.

- Choosing between models is reduced to considering the sign of a nonlinear correlation.
Definition of Linear non-Gaussian Acyclic Model (LiNGAM)

- Given the general, $n$-dimensional SEM

\[ x_i = \sum_{j \neq i} b_{ij} x_j + e_i \text{ for all } i \]

Make the following assumptions:
Definition of Linear non-Gaussian Acyclic Model (LiNGAM)

- Given the general, $n$-dimensional SEM

$$x_i = \sum_{j \neq i} b_{ij} x_j + e_i$$

for all $i$.

Make the following assumptions:

1. The $e_i(t)$ are non-Gaussian, e.g. sparse
   - Crucial departure from classical framework
Definition of Linear non-Gaussian Acyclic Model (LiNGAM)

- Given the general, \( n \)-dimensional SEM
  \[
  x_i = \sum_{j \neq i} b_{ij} x_j + e_i \text{ for all } i
  \]

Make the following assumptions:
1. the \( e_i(t) \) are non-Gaussian, e.g. sparse
   - Crucial departure from classical framework
2. the \( e_i(t) \) are mutually independent
Definition of Linear non-Gaussian Acyclic Model (LiNGAM)

Given the general, $n$-dimensional SEM

$$x_i = \sum_{j \neq i} b_{ij} x_j + e_i \text{ for all } i$$

Make the following assumptions:

1. the $e_i(t)$ are non-Gaussian, e.g. sparse
   - Crucial departure from classical framework
2. the $e_i(t)$ are mutually independent
3. the $b_{ij}$ are acyclic
   - Not completely necessary but simplifies theory
   - Variables can be ordered so that connections only “forward”
   - Could also mean we are analysing the “main directions”
Another viewpoint to importance of non-gaussianity

- A gaussian distribution is completely determined by covariances (and means)
- The number of covariances is \( \approx n^2/2 \) due to symmetry
- So, we cannot solve for the \( \approx n^2 \) connections! ("More variables than equations")
- This is why gaussian methods (PCA, factor analysis, classic SEM) are fundamentally indetermined
- For non-gaussian data, we can use other information than covariances
  - Nonlinear correlations e.g. \( E\{x_i x_j^2\} \), higher-order statistics
As a first approach, we proposed estimation using ICA (Shimizu et al, JMLR, 2006).

Transform

\[ x = Bx + e \Leftrightarrow x = (I - B)^{-1}e \]

Becomes an ICA model

But one complication: ICA does not estimate order of \( e_i \)

- In SEM, \( e_i \) do have a specific order
- Acyclicity allows determination of the right order: Only the right ordering of components allows transformation back to SEM.

This proves identifiability, in contrast to Gaussian case!
Application in functional magnetic resonance imaging (fMRI)

- Specific problems with fMRI when using Granger causality
  - Hemodynamic response functions variable over the cortex (David et al, PLoS Biol, 2008)
  - Granger causality may give very misleading results (S.M. Smith et al, NIMG, 2010)
- Steve Smith et al compared different causal analysis methods with simulated fMRI data.
- Given enough data (250 minutes, TR=3s, 5 variables), LiNGAM worked better than other methods in finding the directionality
- How to make LiNGAM work with less data? Two-variable methods help a lot.
Application of LiNGAM to simulated fMRI

Simulation 1 (5 nodes, 10 minute sessions, TR=3.00s, noise=1.0%, HRFstd=0.5s)
Simulation 2 (10 nodes, 10 minute sessions, TR=3.00s, noise=1.0%, HRFstd=0.5s)
Simulation 3 (15 nodes, 10 minute sessions, TR=3.00s, noise=1.0%, HRFstd=0.5s)
Specific characteristics of EEG and MEG

- In EEG/MEG, connections might be between energies $\sigma_{i,t}^2$ of sources $s_i$
- First, separate sources by ICA, then apply LiNGAM on energies? (Future work)
- Alternatively, generalized autoregressive conditional heteroscedasticity or GARCH (Zhang and Hyvärinen, UAI 2009).
Results of GARCH model on real MEG

Black: positive influence, red: negative influence.
Yellow: manually drawn grouping (Zhang and Hyvärinen, 2009)
Extensions of basic LiNGAM framework

- Latent variables: equivalent to ICA model with more components (Hoyer et al, IJAR 2008)
Extensions of basic LiNGAM framework

- Latent variables: equivalent to ICA model with more components (Hoyer et al, IJAR 2008)
- We can combine instantaneous and lagged effects in the same model (Hyvärinen et al, JMLR, 2010)
Extensions of basic LiNGAM framework

- Latent variables: equivalent to ICA model with more components (Hoyer et al, IJAR 2008)
- We can combine instantaneous and lagged effects in the same model (Hyvärinen et al, JMLR, 2010)
- Cyclicity probably not a major problem in most methods (Lacerda et al 2008)
Extensions of basic LiNGAM framework

- Latent variables: equivalent to ICA model with more components (Hoyer et al, IJAR 2008)
- We can combine instantaneous and lagged effects in the same model (Hyvärinen et al, JMLR, 2010)
- Cyclicity probably not a major problem in most methods (Lacerda et al 2008)
- Group data: individual differences may help in identification (Ramsey, 2011; Shimizu, 2012)
Latent variables: equivalent to ICA model with more components (Hoyer et al, IJAR 2008)

We can combine instantaneous and lagged effects in the same model (Hyvärinen et al, JMLR, 2010)

Cyclicity probably not a major problem in most methods (Lacerda et al 2008)

Group data: individual differences may help in identification (Ramsey, 2011; Shimizu, 2012)

Nonlinear versions (Hoyer et al, 2009, Hyvärinen and Smith, 2013)
Conclusion

- Causal analysis possible using statistics which go beyond correlations
- Structural equation models can be estimated by non-Gaussianity (Shimizu et al, JMLR, 2006)
  - An intuitive approach is likelihood ratios for two variables
  - Alternatively, ICA and re-arrange the coefficients
- Many extensions of basic framework developed
- Applicability to real data, e.g. brain imaging to be determined...