Non-Gaussian Methods for Learning Linear Structural Equation Models

Shohei Shimizu and Yoshinobu Kawahara
Osaka University

Special thanks to Aapo Hyvärinen, Patrik O. Hoyer and Takashi Washio.
Abstract

- Linear structural equation models (linear SEMs) can be used to model data generating processes of variables.

- We review a new approach to learn or estimate linear SEMs.

- The new estimation approach utilizes non-Gaussianity of data for model identification and uniquely estimates much wider variety of models.
Outline

• **Part I.** Overview (70 min.) : Shohei

• **Break** (10 min.)

• **Part II.** Recent advances (40 min): Yoshi
  – Time series
  – Latent confounders
Motivation (1/2)

• Suppose that data \( X \) was randomly generated from either of the following two data generating processes:

\[
\begin{align*}
\text{Model 1:} & & \text{Model 2:} \\
x_1 &= e_1 & x_1 &= b_{12} x_2 + e_1 \\
x_2 &= b_{21} x_1 + e_2 & x_2 &= e_2
\end{align*}
\]

where \( e_1 \) and \( e_2 \) are latent variables (disturbances, errors).

• We want to estimate or identify which model generated the data \( X \) based on the data \( X \) only.
Motivation (2/2)

- We want to identify which model generated the data $X$ based on the data $X$ only.

- If $e_1$ and $e_2$ are Gaussian, it is well known that we cannot identify the data generating process.
  - Models 1 and 2 equally fit data.

- If $e_1$ and $e_2$ are non-Gaussian, an interesting result is obtained: We can identify which of Models 1 and 2 generated the data.

- This tutorial reviews how such non-Gaussian methods work.
Problem formulation
Basic problem setup (1/3)

• Assume that the data generating process of continuous observed variables $x_i$ is graphically represented by a directed acyclic graph (DAG).
  – Acyclicity means that there are no directed cycles.

Example of a directed acyclic graph (DAG):

```
x3  e3
   ↓   ↓
x1  e1
   ↓   ↓
x2  e2
```

$x_3$ is a parent of $x_1$ etc.

Example of a directed cyclic graph:

```
x3  e3
   ↓   ↓
x1  e1
   ↓   ↓
x2  e2
```

$\text{Example of a directed cyclic graph:}$
Basic problem setup (2/3)

• Further assume linear relations of variables $x_i$.
• Then we obtain a linear acyclic SEM (Wright, 1921; Bollen, 1989):

\[ x_i = \sum_{j: \text{parents of } i} b_{ij} x_j + e_i \quad \text{or} \quad x = Bx + e \]

where

– The $e_i$ are continuous latent variables that are not determined inside the model, which we call external influences (disturbances, errors).
– The $e_i$ are of non-zero variance and are independent.
– The ‘path-coefficient’ matrix $B = [b_{ij}]$ corresponds to a DAG.
Example of linear acyclic SEMs

• A three-variable linear acyclic SEM:

\[
\begin{align*}
  x_1 &= 1.5x_3 + e_1 \\
  x_2 &= -1.3x_1 + e_2 \\
  x_3 &= e_3
\end{align*}
\]

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{bmatrix} =
\begin{bmatrix}
  0 & 0 & 1.5 \\
  -1.3 & 0 & 0 \\
  0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3
\end{bmatrix} +
\begin{bmatrix}
  e_1 \\
  e_2 \\
  e_3
\end{bmatrix}
\]

• B corresponds to the data-generating DAG:

- Directed edges:
  - \( b_{ij} = 0 \) \( \iff \) No directed edge from \( x_j \) to \( x_i \)
  - \( b_{ij} \neq 0 \) \( \iff \) A directed edge from \( x_j \) to \( x_i \)
Assumption of acyclicity

- Acyclicity ensures existence of an ordering of variables $x_i$ that makes $B$ lower-triangular with zeros on the diagonal (Bollen, 1989).

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
\end{bmatrix} =
\begin{bmatrix}
  0 & 0 & 1.5 \\
  -1.3 & 0 & 0 \\
  0 & 0 & 0 \\
\end{bmatrix}
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
\end{bmatrix} +
\begin{bmatrix}
  e_1 \\
  e_2 \\
  e_3 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
  x_3 \\
  x_1 \\
  x_2 \\
\end{bmatrix} =
\begin{bmatrix}
  0 & 0 & 0 \\
  1.5 & 0 & 0 \\
  0 & -1.3 & 0 \\
\end{bmatrix}
\begin{bmatrix}
  x_3 \\
  x_1 \\
  x_2 \\
\end{bmatrix} +
\begin{bmatrix}
  e_3 \\
  e_1 \\
  e_2 \\
\end{bmatrix}
\]

The ordering is:

$x_3 < x_1 < x_2$.

$x_3$ may be an ancestor of $x_1$, $x_2$, but not vice versa.
Assumption of independence between external influences

- It implies that there are **no latent confounders** (Spirtes et al. 2000)
  - A latent confounder $f$ is a latent variable that is a parent of more than or equal to two observed variables:

  ![Diagram 1]

  - Such a latent confounder $f$ makes external influences dependent (Part II):

  ![Diagram 2]
Basic problem setup (3/3): Learning linear acyclic SEMs

• Assume that data $X$ is randomly sampled from a linear acyclic SEM (with no latent confounders):

$$x = B x + e$$

• **Goal:** Estimate the path-coefficient matrix $B$ by observing data $X$ only!
  – $B$ corresponds to the data-generating DAG.
Problems:
Identifiability problems of conventional methods
Under what conditions $B$ is identifiable?

- `$B$ is identifiable' $\equiv$ `$B$ is uniquely determined or estimated from $p(x)$'.

- Linear acyclic SEM:
  \[
  x = Bx + e
  \]
  - $B$ and $p(e)$ induce $p(x)$.
  - If $p(x)$ are different for different $B$, then $B$ is uniquely determined.
Conventional estimation principle: Causal Markov condition

• If the data-generating model is a (linear) acyclic SEM, causal Markov condition holds:
  – Each observed variable $x_i$ is independent of its non-descendants in the DAG conditional on its parents (Pearl & Verma, 1991):

$$p(x) = \prod_{i=1}^{p} p(x_i \mid \text{parents of } x_i)$$
Conventional methods based on causal Markov condition

• Methods based on conditional independencies (Spirtes & Glymour, 1991)
  – Many linear acyclic SEMs give a same set of conditional independences and equally fit data.

• Scoring methods based on Gaussianity (Chickering, 2002)
  – Many linear acyclic SEMs give a same Gaussian distribution and equally fit data.

• In many cases, the path-coefficient matrix $B$ is not uniquely determined.
Example

- Two models with Gaussian $e_1$ and $e_2$:

  \begin{align*}
  \text{Model 1:} & \quad x_1 &= e_1 \\
  & \quad x_2 &= 0.8x_1 + e_2
  \\
  \text{Model 2:} & \quad x_1 &= 0.8x_2 + e_1 \\
  & \quad x_2 &= e_2
  \end{align*}

\[ E(e_1) = E(e_2) = 0, \quad \text{var}(x_1) = \text{var}(x_2) = 1 \]

- Both introduce no conditional independence:

\[ \text{cov}(x_1, x_2) = 0.8 \neq 0 \]

- Both induce the same Gaussian distribution:

\[
\begin{bmatrix}
  x_1 \\
  x_2
\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)
\]
A solution:
Non-Gaussian approach
A new direction: Non-Gaussian approach

• Non-Gaussian data in many applications:
  – Neuroinformatics (Hyvarinen et al., 2001); Bioinformatics (Sogawa et al., ICANN2010); Social sciences (Micceri, 1989); Economics (Moneta, Entner, et al., 2010)

• Utilize non-Gaussianity for model identification:
  – Bentler (1983); Mooijaart (1985); Dodge and Rousson (2001)

• The path-coefficient matrix $\mathbf{B}$ is uniquely estimated if $e_i$ are non-Gaussian.
Illustrative example: Gaussian vs non-Gaussian

Model 1:
\[ x_1 = e_1 \]
\[ x_2 = 0.8x_1 + e_2 \]

Model 2:
\[ x_1 = 0.8x_2 + e_1 \]
\[ x_2 = e_2 \]

\[ E(e_1) = E(e_2) = 0, \]
\[ \text{var}(x_1) = \text{var}(x_2) = 1 \]
Linear Non-Gaussian Acyclic Model: LiNGAM
(Shimizu, Hyvarinen, Hoyer & Kerminen, JMLR, 2006)

• Non-Gaussian version of linear acyclic SEM:

\[ x_i = \sum_{j: \text{parents of } i} b_{ij} x_j + e_i \quad \text{or} \quad x = Bx + e \]

where

– The external influence variables \( e_i \) (disturbances, errors) are
  • of non-zero variance.
  • non-Gaussian and mutually independent.
Identifiability of LiNGAM model

• LiNGAM model can be shown to be identifiable.
  – \( B \) is uniquely estimated.

• To see the identifiability, helpful to review independent component analysis (ICA) (Hyvarinen et al., 2001).
Independent Component Analysis (ICA) (Jutten & Herault, 1991; Comon, 1994)

- Observed random vector $\mathbf{x}$ is modeled by

$$x_i = \sum_{j=1}^{p} a_{ij} s_j \quad \text{or} \quad \mathbf{x} = \mathbf{A} \mathbf{s}$$

where

- The mixing matrix $\mathbf{A} = [a_{ij}]$ is square and is of full column rank.
- The latent variables $s_i$ (independent components) are of non-zero variance, non-Gaussian and mutually independent.

- Then, $\mathbf{A}$ is identifiable up to permutation $\mathbf{P}$ and scaling $\mathbf{D}$ of the columns:

$$A_{ica} = APD$$
Estimation of ICA

• Most of estimation methods estimate $W = A^{-1}$:
  
  \[ x = As = W^{-1}s \]

  (Hyvarinen et al., 2001)

• Most of the methods minimize mutual information (or its approximation) of estimated independent components:

  \[ \hat{s} = W_{ica}x \]

• $W$ is estimated up to permutation $P$ and scaling $D$ of the rows:

  \[ W_{ica} = PDW \left( = PDA^{-1} \right) \]

• Consistent and computationally efficient algorithms:
  – Fixed point (FastICA) (Hyvarinen, 99); Gradient-based (Amari, 98)
  – Semiparametric: no specific distributional assumption
Back to LiNGAM model
Identifiability of LiNGAM (1/3): ICA achieves half of identification

- LiNGAM model is ICA.
  - Observed variables $x_i$ are linear combinations of non-Gaussian independent external influences $e_i$:
    \[
    x = Bx + e \iff x = (I - B)^{-1} e
    = Ae = W^{-1} e \quad (W = I - B)
    \]

- ICA gives $W_{ica} = PDW = PD(I - B)$.
  - $P$: unknown permutation matrix
  - $D$: unknown scaling (diagonal) matrix

- Need to determine $P$ and $D$ to identify $B$. 
Identifiability of LiNGAM (2/3): No permutation indeterminacy (1/6)

- ICA gives $W_{ica} = PDW = PD(I - B)$.
  - $P$: permutation matrix; $D$: scaling (diagonal) matrix

- We want to find such a permutation matrix $\overline{P}$ that cancels the permutation $P$, i.e., $\overline{PP} = I$:
  $$\overline{PW}_{ica} = \overline{PPDW} = DW = I$$

- We can show (Shimizu et al., UAI05) (illustrated in the next slides):
  - If $\overline{PP} = I$, i.e., no permutation is made on the rows of $DW$, $\overline{PW}_{ica}$ has no zero in the diagonal (obvious by definition).
  - If $\overline{PP} \neq I$, i.e., any nonidentical permutation is made on the rows of $DW$, $\overline{PW}_{ica}$ has a zero in the diagonal.
Identifiability of LiNGAM (2/3): No permutation indeterminacy (2/6)

• By definition $W = I - B$ has all unities in the diagonal.
  – The diagonal elements of $B$ are all zeros.

• Acyclicity ensures existence of an ordering of variables that makes $B$ lower-triangular, and then $W = I - B$ is also lower-triangular.

• So, WLG, $W$ can be assumed to be lower-triangular:

$$W = \begin{bmatrix} 1 & 0 & 0 \\ \ast & 1 & 0 \\ \ast & \ast & 1 \end{bmatrix}$$

No zeros in the diagonal!
Identifiability of LiNGAM (2/3): No permutation indeterminacy (3/6)

- Premultiplying $\mathbf{W}$ by a scaling (diagonal) matrix $\mathbf{D}$ does NOT change the zero/non-zero pattern of $\mathbf{W}$:

$$
\mathbf{W} = \begin{bmatrix}
1 & 0 & 0 \\
* & 1 & 0 \\
* & * & 1 \\
\end{bmatrix}
\quad \Rightarrow \quad
\mathbf{D}\mathbf{W} = \begin{bmatrix}
d_{11} & 0 & 0 \\
* & d_{22} & 0 \\
* & * & d_{33} \\
\end{bmatrix}
$$

No zeros in the diagonal!
Identifiability of LiNGAM (2/3): No permutation indeterminacy (4/6)

- Any other permutation of the rows of $\mathbf{DW}$ changes the zero/non-zero pattern of $\mathbf{DW}$ and brings zero in the diagonal:

\[
\mathbf{DW} = \begin{bmatrix}
d_{11} & 0 & 0 \\
* & d_{22} & 0 \\
* & * & d_{33}
\end{bmatrix}
\]

\[
\mathbf{P}^{12}\mathbf{DW} = \begin{bmatrix}
* & d_{22} & 0 \\
d_{11} & 0 & 0 \\
* & * & d_{33}
\end{bmatrix}
\]

Exchanging $1^{st}$ and $2^{nd}$ rows

Zero in the diagonal!
Identifiability of LiNGAM (2/3): No permutation indeterminacy (5/6)

• Any other permutation of the rows of $\mathbf{DW}$ changes the zero/non-zero pattern of $\mathbf{DW}$ and brings zero in the diagonal:

$$
\mathbf{D} = \begin{bmatrix}
    d_{11} & 0 & 0 \\
    * & d_{22} & 0 \\
    * & * & d_{33}
\end{bmatrix}
= \begin{bmatrix}
    * & * & d_{33} \\
    * & d_{22} & 0 \\
    d_{11} & 0 & 0
\end{bmatrix}
$$

Exchanging 1st and 3rd rows

Zero in the diagonal!
Identifiability of LiNGAM (2/3): No permutation indeterminacy (6/6)

We can find correct $\bar{P}$ by finding $\bar{P}$ that gives no zero on the diagonal of $\bar{PW}_{ica}$ (Shimizu et al., UAI05).

Thus, we can solve the permutation indeterminacy and obtain:

$$\bar{PW}_{ica} = \bar{PPDW} = DW = D(I - B) = I$$
Identifiability of LiNGAM (3/3): No scaling indeterminacy

- Now we have: $\overline{PW}_{ica} = D(I - B)$

- Then,
  \[ D = \text{diag}(\overline{PW}_{ica}) \]

- Divide each row of $\overline{PW}_{ica}$ by its corresponding diagonal element to get $I - B$, i.e., $B$:
  \[ \text{diag}(\overline{PW}_{ica})^{-1}\overline{PW}_{ica} = D^{-1} D(I - B) = I - B \]
Estimation of LiNGAM model

1. ICA-LiNGAM algorithm
2. DirectLiNGAM algorithm
Two estimation algorithms

• ICA-LiNGAM algorithm
  (Shimizu, Hoyer, Hyvarinen & Kerminen, JMLR, 2006)
• DirectLiNGAM algorithm
  (Shimizu, Hyvarinen, Kawahara & Washio, UAI09)

• Both estimate an ordering of variables that makes the path-coefficient matrix $B$ to be lower-triangular.
  – Acyclicity ensures existence of such an ordering.

\[ x_{\text{perm}} = O \underbrace{B}_{\text{perm}} x_{\text{perm}} + e_{\text{perm}} \]
Once such an ordering is found...

- Many existing (covariance-based) methods can do:
  - Pruning the redundant path-coefficients
    - Sparse methods like weighted lasso (Zou, 2006)
  - Finding significant path-coefficients
    - Testing, bootstrapping (Shimizu et al., 2006; Hyvarinen et al. 2010)

\[
x_{perm} = \begin{bmatrix} \star & 0 & 0 \\ \star & 0 & \star \\ \end{bmatrix} x_{perm} + e_{perm}
\]

A full DAG

- Causal structure: $x_1 \rightarrow x_2 \rightarrow x_3$
1. Outline of ICA-LiNGAM algorithm

(Shimizu, Hoyer, Hyvarinen, & Kerminen, JMLR, 2006)

1. Estimate B by ICA + permutation

2. Pruning

Redundant edges
ICA-LiNGAM algorithm (1/2):
Step 1. Estimation of B

1. Perform ICA (here, FastICA (Hyvarinen, 1999)) to obtain an estimate of

\[ \mathbf{W}_{ica} = \mathbf{PDW} = \mathbf{PD}(\mathbf{I} - \mathbf{B}) \]

2. Find a permutation \( \mathbf{P} \) that makes the diagonal elements of \( \mathbf{P}\hat{\mathbf{W}}_{ica} \) as large (non-zero) as possible in absolute value:

\[
\hat{\mathbf{P}} = \min_{\mathbf{P}} \frac{1}{\left| \left( \mathbf{P}\hat{\mathbf{W}}_{ica} \right)_{ii} \right|}
\]

Hungarian alg. (Kuhn, 1955)

3. Normalize each row of \( \mathbf{P}\hat{\mathbf{W}}_{ica} \), then we get an estimate of \( \mathbf{I-B} \) and \( \hat{\mathbf{B}} \).
ICA-LiNGAM algorithm (2/2): Step 2. Pruning

- Find such an ordering of variables that makes estimated $\mathbf{B}$ be as close to be lower-triangular as possible.
  - Find a permutation matrix $\mathbf{Q}$ that minimizes the sum of the elements in the upper triangular part of permuted $\hat{\mathbf{B}}$:
    \[
    \hat{\mathbf{Q}} = \min_{\mathbf{Q}} \sum_{i \leq j} \left( \mathbf{Q} \hat{\mathbf{B}} \mathbf{Q}^T \right)_{ij}^2
    \]
  - Approximate algorithm for large variables (Hoyer et al., ICA06)
Basic properties of ICA-LiNGAM algorithm

• ICA-LiNGAM algorithm = ICA + permutations
  – Computationally efficient with the help of well-developed ICA techniques.

• Potential problems
  – ICA is an iterative search method:
    • May get stuck in a local optimum if the initial guess or step size is badly chosen.
  – The permutation algorithms are not scale-invariant:
    • May provide different estimates for different scales of variables.
Estimation of LiNGAM model

1. ICA-LiNGAM algorithm
2. DirectLiNGAM algorithm
2. DirectLiNGAM algorithm
(Shimizu, Hyvarinen, Kawahara & Washio, UAI2009)

• Alternative estimation method without ICA
  – Estimates an ordering of variables that makes path-coefficient matrix $B$ to be lower-triangular.

\[
x_{perm} = \begin{bmatrix} O \\ B_{perm} \end{bmatrix} x_{perm} + e_{perm}
\]

• Many existing (covariance-based) methods can do further pruning or finding significant path coefficients (Zou, 2006; Shimizu et al., 2006; Hyvarinen et al. 2010)
Basic idea (1/2):
An exogenous variable can be at the top of a right ordering

• An exogenous variable $x_j$ is a variable with no parents (Bollen, 1989), here $x_3$.
  – The corresponding row of $B$ has all zeros.

• So, an exogenous variable can be at the top of such an ordering that makes $B$ lower-triangular with zeros on the diagonal.

\[
\begin{bmatrix}
  x_3 \\
  x_1 \\
  x_2 \\
\end{bmatrix} = \begin{bmatrix}
  0 & 0 & 0 \\
  1.5 & 0 & 0 \\
  0 & -1.3 & 0 \\
\end{bmatrix} \begin{bmatrix}
  x_3 \\
  x_1 \\
  x_2 \\
\end{bmatrix} + \begin{bmatrix}
  e_3 \\
  e_1 \\
  e_2 \\
\end{bmatrix}
\]
Basic idea (2/2):

**Regress exogenous** $x_3$ **out**

1. Compute the residuals $r_i^{(3)} (i = 1, 2)$ regressing the other variables $x_i (i = 1, 2)$ on exogenous $x_3$:
   - The residuals $r_1^{(3)}$ and $r_2^{(3)}$ form a LiNGAM model.
   - The ordering of the residuals is equivalent to that of corresponding original variables.

\[
\begin{bmatrix}
  x_3 \\
  x_1 \\
  x_2 \\
\end{bmatrix} =
\begin{bmatrix}
  0 & 0 & 0 \\
  1.5 & 0 & 0 \\
  0 & -1.3 & 0 \\
\end{bmatrix}
\begin{bmatrix}
  x_3 \\
  x_1 \\
  x_2 \\
\end{bmatrix} +
\begin{bmatrix}
  e_3 \\
  e_1 \\
  e_2 \\
\end{bmatrix}
\Rightarrow
\begin{bmatrix}
  r_1^{(3)} \\
  r_2^{(3)} \\
\end{bmatrix} =
\begin{bmatrix}
  0 & 0 \\
  -1.3 & 0 \\
\end{bmatrix}
\begin{bmatrix}
  r_1^{(3)} \\
  r_2^{(3)} \\
\end{bmatrix} +
\begin{bmatrix}
  e_1 \\
  e_2 \\
\end{bmatrix}
\]

2. Exogenous $r_1^{(3)}$ implies ` $x_1$ can be at the second top'.
Outline of DirectLiNGAM

• **Iteratively** find exogenous variables until all the variables are ordered:

  1. Find an exogenous variable $x_3$.
     - Put $x_3$ at the top of the ordering.
     - Regress $x_3$ out.

  2. Find an exogenous residual, here $r_1^{(3)}$.
     - Put $x_1$ at the second top of the ordering.
     - Regress $r_1^{(3)}$ out.

  3. Put $x_2$ at the third top of the ordering and terminate.
     The estimated ordering is $x_3 < x_1 < x_2$.

![Diagram](image)
Identification of an exogenous variable (two variable cases)

i) $x_1 (= e_1)$ is exogenous.

\[
x_1 = e_1 \\
x_2 = b_{21} x_1 + e_2 \quad (b_{21} \neq 0)
\]

Regressing $x_2$ on $x_1$,

\[
r_{2(1)} = x_2 - \frac{\text{cov}(x_2, x_1)}{\text{var}(x_1)} x_1 \\
= x_2 - b_{21} x_1 \\
= e_2
\]

$x_1$ and $r_{2(1)}$ are independent.

ii) $x_1$ is NOT exogenous.

\[
x_1 = b_{12} x_2 + e_1 \quad (b_{12} \neq 0) \\
x_2 = e_2
\]

Regressing $x_2$ on $x_1$,

\[
r_{2(1)} = x_2 - \frac{\text{cov}(x_2, x_1)}{\text{var}(x_1)} x_1 \\
= \left\{ 1 - \frac{b_{12} \text{cov}(x_2, x_1)}{\text{var}(x_1)} \right\} x_2 - \frac{b_{12} \text{var}(x_2)}{\text{var}(x_1)} e_1
\]

$x_1$ and $r_{2(1)}$ are NOT independent.
Need to use Darmois-Skitovitch’ theorem (Darmois, 1953; Skitovitch, 1953)

Darmois-Skitovitch’ theorem:

Define two variables $x_1$ and $x_2$ as

$$x_1 = \sum_{j=1}^{p} a_{1j} e_j, \quad x_2 = \sum_{j=1}^{p} a_{2j} e_j$$

where $e_j$ are independent random variables.

If there exists a non-Gaussian $e_i$ for which $a_{1i}a_{2i} \neq 0$, $x_1$ and $x_2$ are dependent.

ii) $x_1$ is NOT exogenous.

$$x_1 = b_{12} x_2 + 1 \cdot e_1 \quad (b_{12} \neq 0)$$

$$x_2 = e_2$$

Regressing $x_1$ on $x_2$,

$$r_2^{(1)} = x_2 - \frac{\text{cov}(x_2, x_1)}{\text{var}(x_1)} x_1$$

$$= \left\{ 1 - \frac{b_{12} \text{cov}(x_2, x_1)}{\text{var}(x_1)} \right\} x_2 - \frac{b_{12} \text{var}(x_2)}{\text{var}(x_1)} e_1$$

$x_1$ and $r_2^{(1)}$ are NOT independent.
Identification of an exogenous variable (p variable cases)

• Lemma 1: $x_j$ and its residual $r_i^{(j)} = x_i - \frac{\text{cov}(x_i, x_j)}{\text{var}(x_j)} x_j$

are independent for all $i \neq j \iff x_j$ is exogenous

• In practice, we can identify an exogenous variable by finding a variable that is most independent of its residuals.
Independence measures

• Evaluate independence between a variable and a residual by a nonlinear correlation:

\[ \left| \text{corr}\{x_j, g(r_i^{(j)})\} \right| \quad (g = \tanh) \]

• Taking the sum over all the residuals, we get:

\[ T = \sum_{i \neq j} \left| \text{corr}\{x_j, g(r_i^{(j)})\} + \text{corr}\{g(x_j), r_i^{(j)}\} \right| \]

• Can use more sophisticated measures as well (Bach & Jordan, 2002; Gretton et al., 2005; Kraskov et al., 2004).
  – Kernel-based independence measure (Bach & Jordan, 2002) often gives more accurate estimates (Sogawa et al., IJCNN10).
Important properties of DirectLiNGAM

• DirectLiNGAM repeats:
  – Least squares simple linear regression
  – Evaluation of pairwise independence between each variable and its residuals.

• No algorithmic parameters like stepsize, initial guesses, convergence criteria.

• Guaranteed convergence to the right solution in a fixed number of steps (the number of variables) if the data strictly follows the model.
Estimation of LiNGAM model: Summary (1/2)

• Two estimation algorithms:
  – ICA-LiNGAM: Estimation using ICA
    • Pros. Fast
    • Cons. Possible local optimum; Not scale-invariant
  – DirectLiNGAM: Alternative estimation without ICA
    • Pros. Guaranteed convergence; Scale-invariant
    • Cons. Less fast
  – Cf. Neither needs faithfulness (Shimizu et al., JMLR, 2006; Hoyer, personal comm., July, 2010).
Estimation of LiNGAM model: Summary (2/2)

- Experimental comparison of the two algorithms: (Sogawa et al., IJCNN2010)

- **Sample size**: Both need at least 1000 sample size for more than 10 variables.

- **Scalability**: Both can analyze 100 variables. The performances depend on the sample size etc., of course!

- **Scale invariance**: ICA-LiNGAM is less robust for changing scales of variables.

- **Local optima?**:
  - For less than 10 variables, ICA-LiNGAM often a bit better.
  - For more than 10 variables, DirectLiNGAM often better perhaps because the problem of local optima might become more serious?
Testing and Reliability evaluation
Testing testable assumptions

- Non-Gaussianity of external influences $e_i$:
  - Gaussianity tests

- Could detect violations of some assumptions:
  - Local test
    - Independence of external influences $e_i$
    - Conditional independencies between observed variables $x_i$ (causal Markov condition)
    - Linearity
  - Overall fit of the model assumptions
    - Chi-square test using 3rd and/or 4th-order moments (Shimizu & Kano, 2008)
    - Still under development
Reliability evaluation

• Need to evaluate statistical reliability of LiNGAM results:
  – Sample fluctuations
  – Smaller non-Gaussianity makes the model closer to be NOT identifiable.

• Reliability evaluation by bootstrapping:
  (Komatsu, Shimizu & Shimodaira, ICANN2010; Hyvarinen et al., JMLR, 2010)
  – If either the sample size is small or the magnitude of non-Gaussianity is small, LiNGAM would give very different results for bootstrap samples.
Extensions
Extensions (a partial list)

• Relaxing the assumptions of LiNGAM model:
  – Acyclic $\rightarrow$ Cyclic (Lacerda et al., UAI2008)
  – Single homogenous population $\rightarrow$ heterogeneous population (Shimizu et al., ICONIP2007)
  – i.i.d. sampling $\rightarrow$ time structures (Part II.) (Hyvarinen et al., JMLR, 2010; Kawahara, S & Washio, 2010)
  – No latent confounders $\rightarrow$ Allow latents (Part II.) (Hoyer et al., IJAR, 08; Kawahara, Bollen et al., 2010)
  – Linear $\rightarrow$ Non-linear (Hoyer et al., NIPS08; Zhang & Hyvarinen, UAI09; Tilmann, Gretton & Spirtes, NIPS09)
Application areas so far
Non-Gaussian SEMs have been applied to...

- **Neuroinformatics**
  - Brain connectivity analysis (Hyvarinen et al., JMLR, 2010)

- **Bioinformatics**
  - Gene network estimation (Sogawa et al., ICANN2010)

- **Economics** (Wan & Tan, 2009; Moneta, Entner, Hoyer & Coad, 2010)

- **Genetics** (Ozaki & Ando, 2009)

- **Environmental sciences** (Niyogi et al., 2010)

- **Physics** (Kawahara, Shimizu & Washio, 2010)

- **Sociology** (Kawahara, Bollen, Shimizu & Washio, 2010)
Final summary of Part I

- Use of non-Gaussianity in linear SEMs is useful for model identification.
- Non-Gaussian data is encountered in many applications.
- The non-Gaussian approach can be a good option.
- Links to codes and papers: http://homepage.mac.com/shoheishimizu/lingampapers.html
FAQs
Q. My data is Gaussian. LiNGAM will not be useful.

- A. You’re right. Try Gaussian methods.

- Comment: Hoyer et al. (UAI2008) showed: `To what extent one can identify the model for a mixture of Gaussian and non-Gaussian external influence variables’.
Q. I applied LiNGAM, but the result is not reasonable to background knowledge.

• A. You might first want to check:
  – Some model assumptions might be violated.  
    → Try other extensions of LiNGAM or non-parametric methods PC or FCI etc. (Spirtes et al., 2000).
  – Small sample size or small non-Gaussianity  
    → Try bootstrap to see if the result is reliable.
  – Background knowledge might be wrong.
Q. Relation to causal Markov condition?

- A. The following 3 estimation principles are equivalent (Zhang & Hyvarinen, ECML09; Hyvarinen et al., JMLR, 2010):
  1. Maximize independence between external influences $e_i$.
  2. Minimize the sum of entropies of external influences $e_i$.
  3. Causal Markov condition (Each variable is independent of its non-descendants in the DAG conditional on its parents) AND maximization of independence between the parents of each variable and its corresponding external influences $e_i$. 
Q. I am a psychometrician and am more interested in latent factors.

- A. Shimizu, Hoyer, and Hyvarinen. (2009) proposes LiNGAM for latent factors:

\[ f = Bf + d \quad \text{-- LiNGAM for latent factors} \]

\[ x = Gf + e \quad \text{-- Measurement model} \]
Others

• Q. Prior knowledge?
  – It is possible to incorporate prior knowledge. The accuracy of DirectLiNGAM is often greatly improved even if the amount of prior knowledge is not so large (Inazumi et al., LVA/ICA2010).

• Q. Sparse LiNGAM?
  – Zhang et al. (ICA09) and Hyvarinen et al. (JMLR, 2010).
  – ICA + adaptive Lasso (Zou, 2006).

• Q. Bayesian approach?
  – Hoyer and Hyttinen (UAI09); Henao and Winther. (NIPS09).

• Q. The idea can be applied to discrete variables?
  – One proposal by Peters et al. (AISTATS2010).
  – Comment: if your discrete variables are close to be continuous, e.g., ordinal scales with many points, LiNGAM might work.
Q. Nonlinear extensions?

• A. Several nonlinear SEMs have been proposed:
  – DAG; No latent confounders.

1. \[ x_i = \sum_j f_{ij}(\text{parent } j \text{ of } x_i) + e_i \]  -- Imoto et al. (2002)
2. \[ x_i = f_i(\text{parents of } x_i) + e_i \]  -- Hoyer et al. (NIPS08)
3. \[ x_i = f_{i,2}^{-1}(f_{i,1}(\text{parents of } x_i) + e_i) \]  -- Zhang et al. (UAI09)

• For two variable cases, unique identification possible except several combinations of nonlinearities and distributions (Hoyer et al., NIPS08; Zhang & Hyvarinen, UAI09).
Nonlinear extensions (continued)

- Proposals to aim at computational efficiency (Mooij et al., ICML09; Tilmann et al., NIPS09; Zhang & Hyvarinen, ECML09; UAI09).

- **Pros:**
  - Nonlinear models are more general than linear models.

- **Cons:**
  - Computationally demanding.
    - Current: at most 7 or 8 variables.
    - Perhaps, assumption of Gaussian external influences might help: Imoto et al. (2002) analyzes 100 variables.
  - More difficult to allow other possible violations of LiNGAM assumptions, latent confounders etc.
Q. My data follows neither such linear SEMs nor such nonlinear SEMs as you have talked.

• A. Try non-parametric methods, e.g.,
  – DAG: PC (Spirtes & Glymour, 1991)
  – DAG with latent confounders: FCI (Spirtes et al., 1995).

\[ x_i = f_i(\text{parents of } x_i, e_i) \]

• Probably you get an (probably large) class of equivalence models rather than a single model, but that would be the best you currently can.
Q. Deterministic relations?

• A. LiNGAM is not applicable.

• See Daniusis et al. (UAI2010) for a method to analyze deterministic relations.
Extra slides
Very brief review of Linear SEMs and causality

See Pearl (2000) for details.
We can give a causal interpretation to each path-coefficient $b_{ij}$ (Pearl, 2000).

- Example:
  
  $x_1 = e_1$

  $x_2 = b_{21}x_1 + e_2$

  $x_3 = b_{32}x_2 + e_3$

- Causal interpretation of $b_{32}$:

  If you change $x_2$ from a constant $c_0$ to a constant $c_1$, then $E(x_3)$ will change by $b_{32}(c_1 - c_0)$. 

`Change $x_2$ from $c_0$ to $c_1$`

- It means `Fix $x_2$ at $c_0$ regardless of the variables that previously determines $x_2$ and change the constant from $c_0$ to $c_1` (Pearl, 2000).

- `Fix $x_2$ at $c_0$ regardless of the variables that determines $x_2$` is denoted by $do(x_2 = c_0)$.
  - In SEM, `Replace the function determining $x_2$ with a constant $c_0`:

\[ x_1 = e_1 \]
\[ x_2 = b_{21}x_1 + e_2 \]
\[ x_3 = b_{32}x_2 + e_3 \]

\[ x_1 = e_1 \]
\[ x_2 = c_0 \]
\[ x_3 = b_{32}x_2 + e_3 \]

(Assumption of invariance: the other functions don’t change)
Average causal effect:
Definition (Rubin, 1974; Pearl, 2000)

• Average causal effect of $x_2$ on $x_3$ when changing $x_2$ from $c_0$ to $c_1$:

$$E(x_3 \mid do(x_2 = c_1)) - E(x_3 \mid do(x_2 = c_0)) = b_{32}(c_1 - c_0)$$

  – Computed based on the mutilated models with $do(x_2 = c_1)$ or $do(x_2 = c_0)$.

• Average causal effects can be computed based on estimated path-coefficient matrix $B$. 
Cf. Conducting experiments with random assignment is a method to estimate average causal effects

- Example: a drug trial
- Random assignment of subjects into two groups fixes $x_2$ to $c_1$ or $c_0$ regardless of the variables that determines $x_2$:
  - Experimental group with $x_2 = c_1$ (take the drug)
  - Control group with $x_2 = c_0$ (placebo)

\[
E(x_3 \mid do(x_2 = c_1)) - E(x_3 \mid do(x_2 = c_0))
\]

$E(x_3)$ for the exp. group - $E(x_3)$ for the control group