

IMPS 2008, Durham, NH

# Linear Non-Gaussian Structural Equation Models

Shohei Shimizu, Patrik Hoyer and Aapo Hyvarinen



Osaka University, Japan  
University of Helsinki, Finland

# Abstract

- Linear Structural Equation Modeling (linear SEM)
  - Analyzes causal relations
- *Covariance*-based SEM
  - Uses **covariance structure** alone for model identification
  - A number of **indistinguishable** models
- Linear *non-Gaussian* SEM
  - Uses **non-Gaussian structures** for model identification
  - Makes many models **distinguishable**

# SEM and causal analysis

- SEM is often used for causal analysis based on non-experimental data
- **Assumption:** the data generating process is represented by a SEM model
- If the assumption is reasonable, SEM provides causal information

# Limitations of covariance-based SEM

- Covariance-based SEM cannot distinguish between many models
- Example



# Linear **non-Gaussian** SEM

- Many observed data are considerably non-Gaussian (Micceri, 1989; Hyvarinen et al. 2001)
- Non-Gaussian structures of data are useful (Bentler 1983; Mooijaart 1985)
- Non-Gaussianity distinguish between the two models (Shimizu et al. 2006) :



# Independent component analysis (ICA)

- Observed random vector  $x$  is modeled as

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

- $s_i$  are independent and **non-Gaussian**
  - Zero means and unit variances
- $A$  is a constant matrix
  - Typically square, # variables = # independent components
- Identifiable up to permutation of the columns  
(Mooijaart 1985; Comon, 1994)

# ICA estimation

- An alternative expression of ICA ( $\mathbf{x}=\mathbf{A}\mathbf{s}$ ):

$$\mathbf{s} = \tilde{\mathbf{W}}\mathbf{x},$$

where  $\tilde{\mathbf{W}} = \mathbf{A}^{-1}$  called a recovering matrix

# ICA estimation

- An alternative expression of ICA ( $\mathbf{x}=\mathbf{A}\mathbf{s}$ ):

$$\mathbf{s} = \tilde{\mathbf{W}}\mathbf{x},$$

where  $\tilde{\mathbf{W}} = \mathbf{A}^{-1}$  called a recovering matrix

- Find such  $\mathbf{W}$  that maximizes independence of components of  $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$ 
  - Many proposals (Hyvarinen et al. 2001)

# ICA estimation

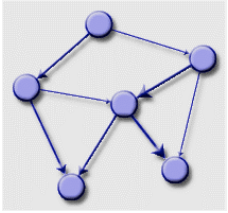
- An alternative expression of ICA ( $\mathbf{x}=\mathbf{A}\mathbf{s}$ ):

$$\mathbf{s} = \tilde{\mathbf{W}}\mathbf{x},$$

where  $\tilde{\mathbf{W}} = \mathbf{A}^{-1}$  called a recovering matrix

- Find such  $\mathbf{W}$  that maximizes independence of components of  $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$ 
  - Many proposals (Hyvarinen et al. 2001)
- $\tilde{\mathbf{W}}$  is estimated up to permutation of the rows:

$$\mathbf{W} = \mathbf{P}\tilde{\mathbf{W}}$$



# Discovery of linear non-Gaussian acyclic models

Shimizu, Hoyer, Hyvarinen and Kerminen (2006)

# Linear **non-Gaussian** acyclic model (LiNGAM)

- Directed acyclic graphs (DAG)
  - $x_i$  can be arranged in a order  $k(i)$
- Assumptions:
  - Linearity
  - External influences  $e_i$  are **independent**
  - and are **non-Gaussian**

$$x_i = \sum_{k(j) < k(i)} b_{ij} x_j + e_i \quad \text{or} \quad \mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

# Goal

- We know
  - Data X is generated by  $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$
- We do **NOT** know
  - Path coefficients:  $b_{ij}$
  - Orders  $k(i)$
  - External influences:  $e_i$
- What we observe is data X only
- Goal
  - Estimate B and  $k(i)$  using data X only!

# Key idea

- First, relate LiNGAM with ICA as follows:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

$$\Rightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e} = \mathbf{A}\mathbf{e} \quad - \text{ICA!}$$

# Key idea

- First, relate LiNGAM with ICA as follows:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

$$\Rightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e} = \mathbf{A}\mathbf{e} \quad - \text{ICA!}$$

$$\text{equivalently } \mathbf{e} = (\mathbf{I} - \mathbf{B})\mathbf{x} = \tilde{\mathbf{W}}\mathbf{x}$$

# Key idea

- First, relate LiNGAM with ICA as follows:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

$$\Rightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e} = \mathbf{A}\mathbf{e} \quad - \text{ICA!}$$

$$\text{equivalently } \mathbf{e} = (\mathbf{I} - \mathbf{B})\mathbf{x} = \tilde{\mathbf{W}}\mathbf{x}$$

- Due to the permutation indeterminacy, ICA gives:

$$\mathbf{W} = \mathbf{P}\tilde{\mathbf{W}}$$

# Key idea

- First, relate LiNGAM with ICA as follows:

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

$$\Rightarrow \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e} = \mathbf{A}\mathbf{e} \quad \text{- ICA!}$$

$$\text{equivalently } \mathbf{e} = (\mathbf{I} - \mathbf{B})\mathbf{x} = \tilde{\mathbf{W}}\mathbf{x}$$

- Due to the permutation indeterminacy, ICA gives:

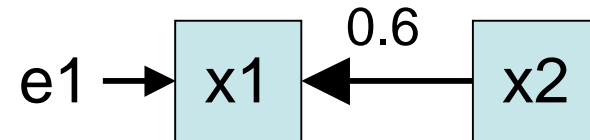
$$\mathbf{W} = \mathbf{P}\tilde{\mathbf{W}}$$

- Can find the correct  $\mathbf{P}$ 
  - The correct permutation is the only one that has **no zeros in the diagonal**

# Illustrative example

- Consider the model:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0.6 \\ 0 & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$



- Goal
  - Estimate the path direction between  $x_1$  and  $x_2$  observing only  $x_1$  and  $x_2$

# Perform ICA

- Relation of the LiNGAM model with ICA:

$$\mathbf{e} = \tilde{\mathbf{W}}\mathbf{x} \quad \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -0.6 \\ 0 & 1 \end{bmatrix}}_{\tilde{\mathbf{W}}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

# Perform ICA

- Relation of the LiNGAM model with ICA:

$$\mathbf{e} = \tilde{\mathbf{W}}\mathbf{x} \quad \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -0.6 \\ 0 & 1 \end{bmatrix}}_{\tilde{\mathbf{W}}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Due to the permutation indeterminacy, ICA might give:

$$\mathbf{W} (= \mathbf{P}\tilde{\mathbf{W}}) = \begin{bmatrix} 0 & 1 \\ 1 & -0.6 \end{bmatrix}$$

# Find the correct P

- Find a permutation of the rows of W so that it has **no zeros in the diagonal**
- In the example...

$$\begin{bmatrix} e_2 \\ e_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & -0.6 \end{bmatrix}}_{\mathbf{W}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Permute the rows



# Find the correct P

- Find a permutation of the rows of  $W$  so that it has **no zeros in the diagonal**
- In the example...

$$\begin{array}{c} \begin{bmatrix} e_2 \\ e_1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -0.6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ \underbrace{\hspace{10em}} \\ \mathbf{W} \end{array}$$

Permute the rows



# Find the correct P

- Find a permutation of the rows of  $W$  so that it has **no zeros in the diagonal**
- In the example...

$$\begin{bmatrix} e_2 \\ e_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & -0.6 \end{bmatrix}}_W \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Permute the rows



$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & -0.6 \\ 0 & 1 \end{bmatrix}}_{\tilde{W}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

# Find the correct P

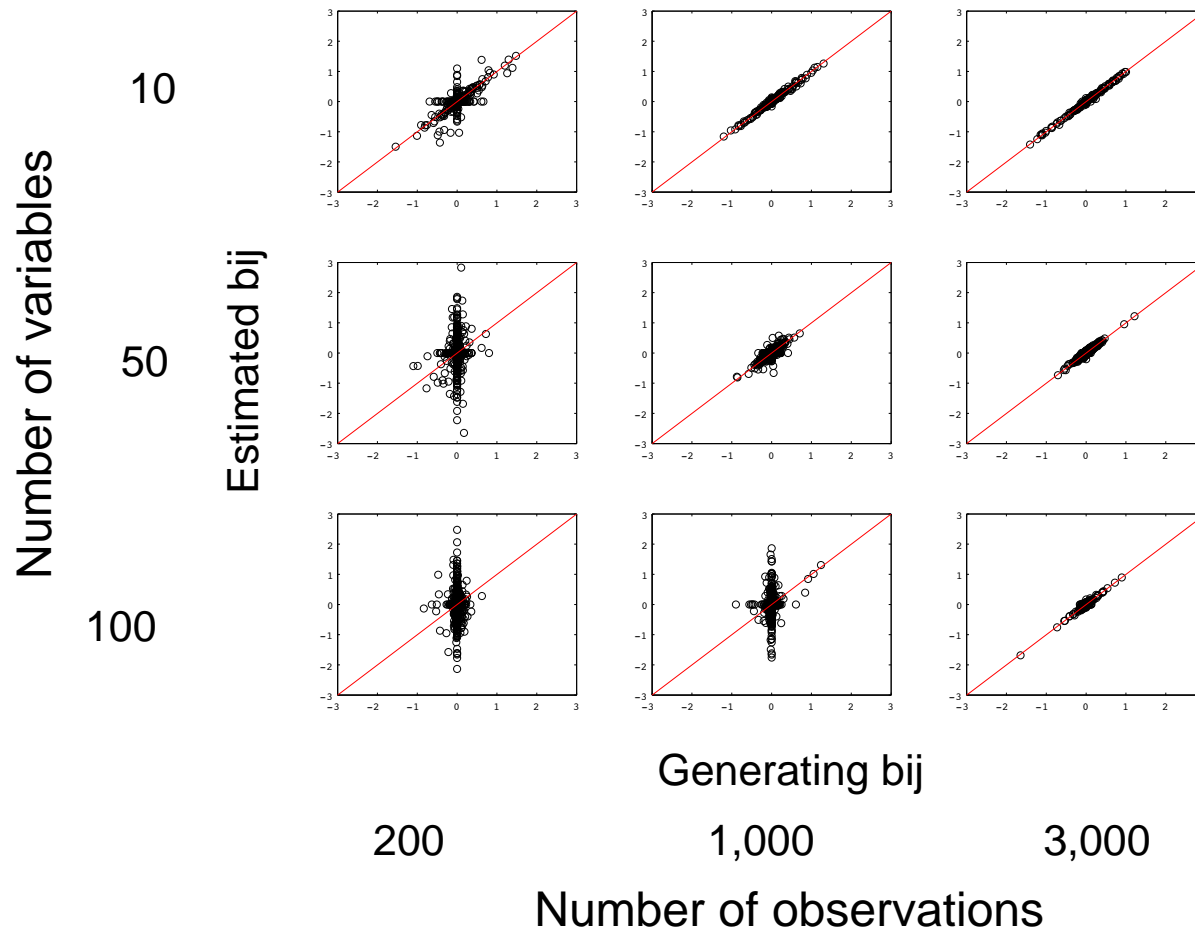
- In practice,

$$\hat{\mathbf{P}} = \max_{\mathbf{P}} \frac{1}{\left| \left( \mathbf{P}^T \mathbf{W} \right)_i \right|}$$

- Heavily penalizes small absolute values in the diagonal

# Simulations: Estimation of B

- Both super- and sub-Gaussian external influences tested
- 5 datasets created for each scatterplot
- B randomly generated at each trial



# Prune B (1)

- In practice, due to estimation errors, we would get:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0.65 \\ -0.05 & 0 \end{bmatrix}}_{\mathbf{B}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

- Need to find which path coefficients are actually zeros

# Find a permutation that gives a lower triangular matrix

- The LiNGAM model is acyclic
  - The matrix  $B$  can be permuted to be lower triangular for some permutation of variables (Bollen, 1989)

# Find a permutation that gives a lower triangular matrix

- The LiNGAM model is acyclic
  - The matrix  $B$  can be permuted to be lower triangular for some permutation of variables (Bollen, 1989)
- First, find a *simultaneous* permutation of rows and columns of  $B$  that gives a *lower-triangular*  $B$

# Find a permutation that gives a lower triangular matrix

- The LiNGAM model is acyclic
  - The matrix B can be permuted to be lower triangular for some permutation of variables (Bollen, 1989)
- First, find a *simultaneous* permutation of rows and columns of B that gives a *lower-triangular* B
- In practice, find a permutation matrix Q that minimizes the sum of the elements in its *upper triangular* part:

$$\hat{Q} = \min_Q \sum_{i \leq j} (QBQ^T)_{ij}$$

# Get a lower-triangular B

- Applying such a simultaneous permutation of the rows and columns,
- we get a permuted B that is as lower-triangular as possible

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0.65 \\ -0.05 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad \Rightarrow$$

**B**

# Get a lower-triangular B

- Applying such a simultaneous permutation of the rows and columns,
- we get a permuted B that is as lower-triangular as possible

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0.65 \\ -0.05 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & \underline{-0.05} \\ 0.65 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_1 \end{bmatrix}$$

**B** **QBQ<sup>T</sup>**

# Get a lower-triangular B

- Applying such a simultaneous permutation of the rows and columns,
- we get a permuted B that is as lower-triangular as possible
- Set the upper-triangular elements to be zeros

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0.65 \\ -0.05 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & \underline{-0.05} \\ 0.65 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_1 \end{bmatrix}$$

**B** **QBQ<sup>T</sup>**

# Get a lower-triangular B

- Applying such a simultaneous permutation of the rows and columns,
- we get a permuted B that is as lower-triangular as possible
- Set the upper-triangular elements to be zeros

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0.65 \\ -0.05 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & \underline{0} \\ 0.65 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_1 \end{bmatrix}$$

**B** **QBQ<sup>T</sup>**

# Pruning B (2)

- Once we get a lower-triangular B, the model is identifiable using covariance-based SEM

$$\begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0.65 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_1 \end{bmatrix}$$

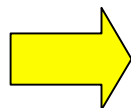
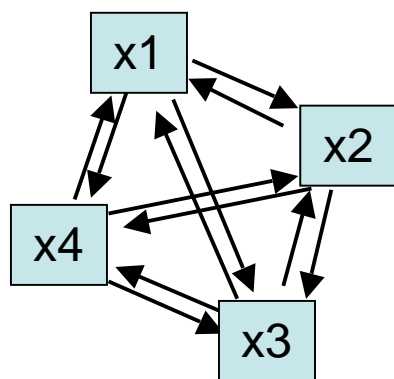
- Many existing methods can be used for pruning the remaining path coefficients
  - Wald test, Bootstrapping, Model fit
  - Lasso-type estimators (Tibshirani 1996; Zou, 2006) etc.

# To summarize the procedure...

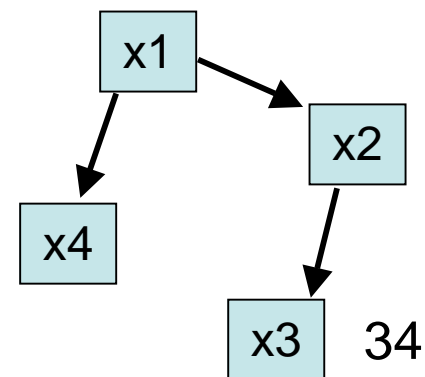
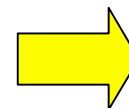
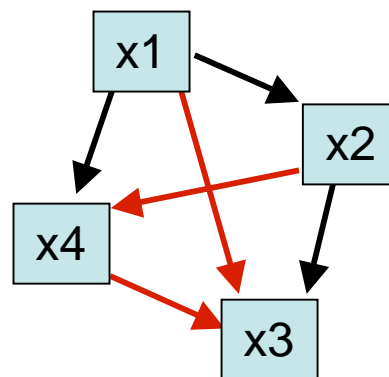
1. Estimate B
  - ICA + finding the correct row permutation
2. Prune estimated B
  1. Find a row-and-column permutation that makes estimated B lower triangular
  2. Prune remaining paths using a covariance-based method

---

## 1. Estimate B



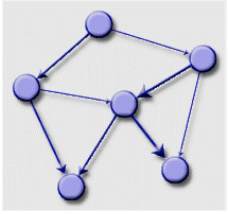
## 2. Prune estimated B



34

# Summary of the regular LiNGAM

- A linear acyclic model is identifiable based on non-Gaussianity
- ICA-based estimation works well
  - Confidence intervals (Konya et al., in progress)
- Better pruning methods might be developed
  - Imposing sparseness in the ICA stage (Zhang & Chang, 2006; Hayashi et al. in progress) like LASSO (Tibshirani 1996)



Some extensions

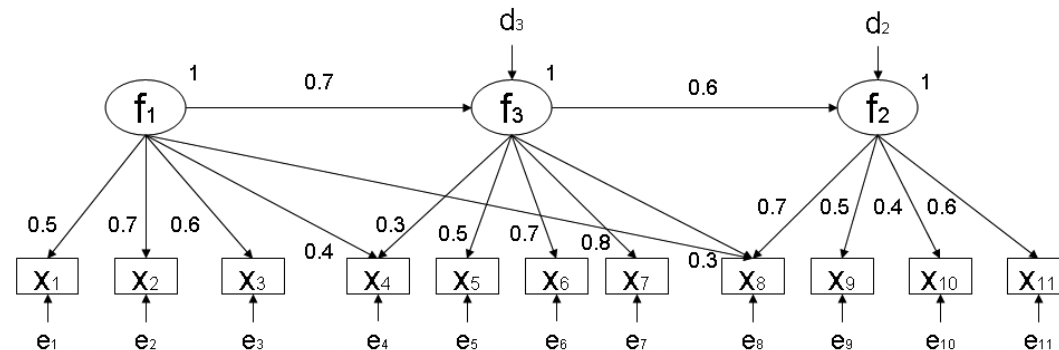
# Latent factors (Shimizu et al., 2007)

- A non-Gaussian multiple indicator model:

$$\mathbf{f} = \mathbf{Bf} + \mathbf{d}$$

$$\mathbf{x} = \mathbf{Gf} + \mathbf{e}$$

- Suppose that  $\mathbf{G}$  is identified, then  $\mathbf{B}$  is identified
  - Could identify  $\mathbf{G}$  in a data driven way using a tetrad-constraint-based method (Silva et al., 2006)



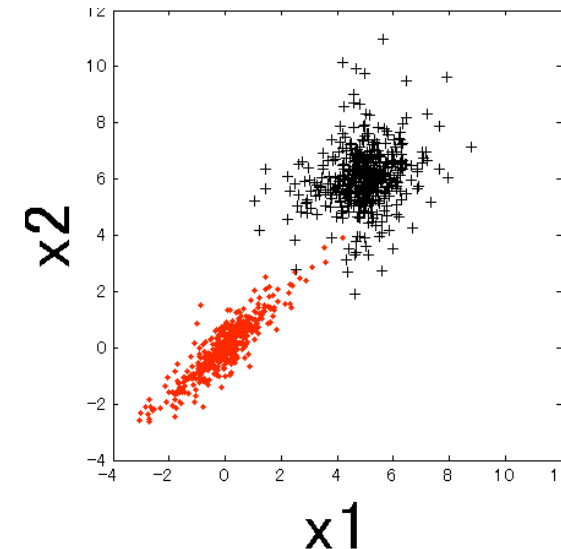
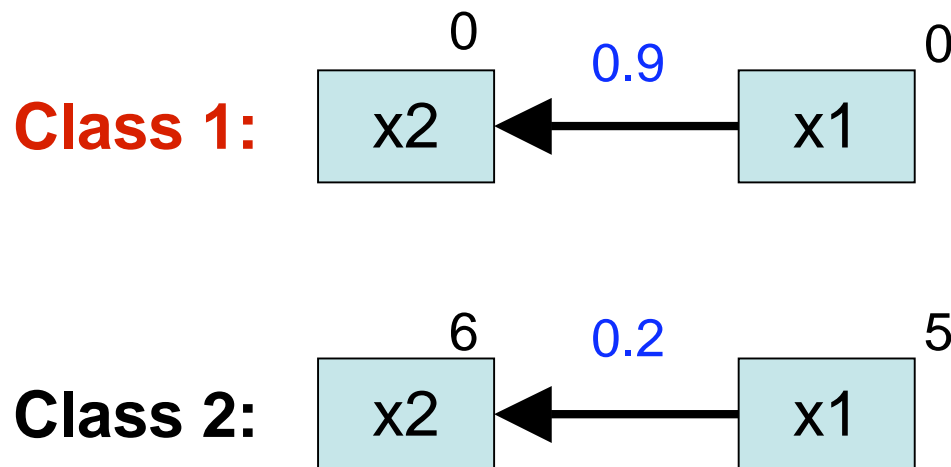
# Latent classes

(Shimizu & Hyvarinen, 2008)

- LiNGAM model for each class  $q$ :

$$\mathbf{x} = \mathbf{B}_q \mathbf{x} + (\mathbf{I} - \mathbf{B}_q) \mathbf{i}_q + \mathbf{e}_q \Rightarrow \mathbf{x} = \mathbf{i}_q + \mathbf{A}_q \mathbf{e}_q \quad - \text{ICA!}$$

- ICA mixtures (Lee et al., 2000; Mollah et al., 2006)



# Unobserved confounders

(Hoyer et al., in press)

- Can identify and distinguish between more models

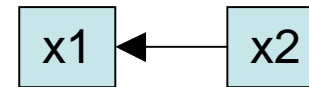
1.



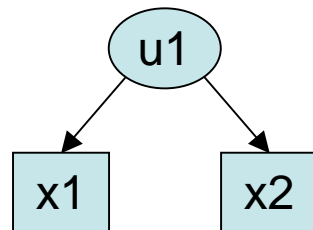
2.



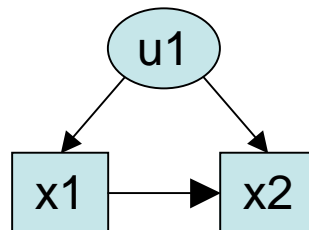
3.



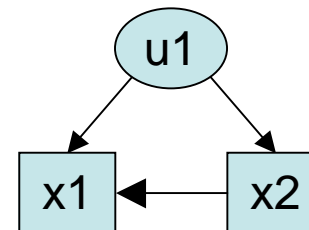
4.



5.



6.



# Time structures (Hyvarinen et al., 2008)

- Combining LiNGAM and autoregressive model:

$$\mathbf{x}(t) = \sum_{\tau=0}^k \mathbf{B}_{\tau} \mathbf{x}(t - \tau) + \mathbf{e}(t)$$

- In econometrics: Structural vector autoregression  
(Swanson & Granger, 1997)

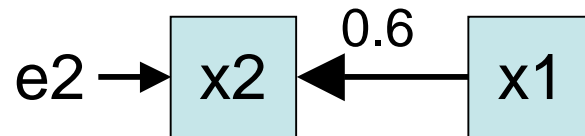
- Changes ordinary AR coefficients based on instantaneous effects:

$$\mathbf{B}_{\tau} = (\mathbf{I} - \mathbf{B}_0) \mathbf{M}_{\tau} \text{ for } \tau > 0 \quad (\mathbf{M}_{\tau} : \text{AR matrix})$$

# Some variables are Gaussian

(Hoyer et al., 2008)

- Consider the model:



- Can identify the path direction
  - if either of  $x_1$  or  $e_2$  is non-Gaussian
- In general, there exist several **equivalent models** that *entail the same distribution* if some are Gaussian

# Some other extensions

- **Cyclic models** (Lacerda et al., 2008)
  - Fewer equivalent models than covariance-based approach
- **Nonlinearity** (Zhang & Chan, 2007; Sun et al., 2007)
- **Model fit statistics are under development**
  - Non-Gaussian structures

# Conclusion

- Use of non-Gaussianity in SEM is useful for model identification
- Many observed data are considerably non-Gaussian
- The non-Gaussian approach can be a good option

- Most of our papers and Matlab/Octave code are available on our webpages
- Google will find us!