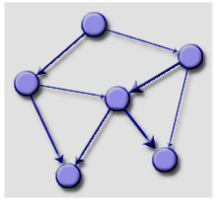


NIPS 2006 Workshop
Causality and Feature selection

Discovery of linear acyclic models in the presence of latent classes using ICA mixtures

Shohei Shimizu and Aapo Hyvarinen

Institute of Statistical Mathematics, Japan
University of Helsinki, Finland

Statistical causal inference for empirical research

- An effective way is to conduct an experiment with random assignment (Holland 1989, Rubin 1974)
- Many situations where it is costly or esthetically difficult to conduct experiments
- Need to develop useful methods to find likely causal models from non-experimental data
- We have proposed a new method for continuous-valued data using independent component analysis called LiNGAM
- We extend the method to cases where latent classes are present (a nonlinear extension)

Independent Component Analysis

(Comon 1994; Hyvarinen et al., 2001)

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

- A variant of factor analysis
- A is unknown. Typically, A is square
- s_i (continuous variables) are assumed to be non-Gaussian and independent
- Estimable including the rotation (Comon, 1994)

Linear acyclic models

(Bollen, 1989; Pearl, 2000; Spirtes et al., 2000)

- Continuous variables
- Assumptions
  1. Directed acyclic graph (DAG)
  2. Linearity
  3. Disturbances (errors) are independent (no hidden confounders)

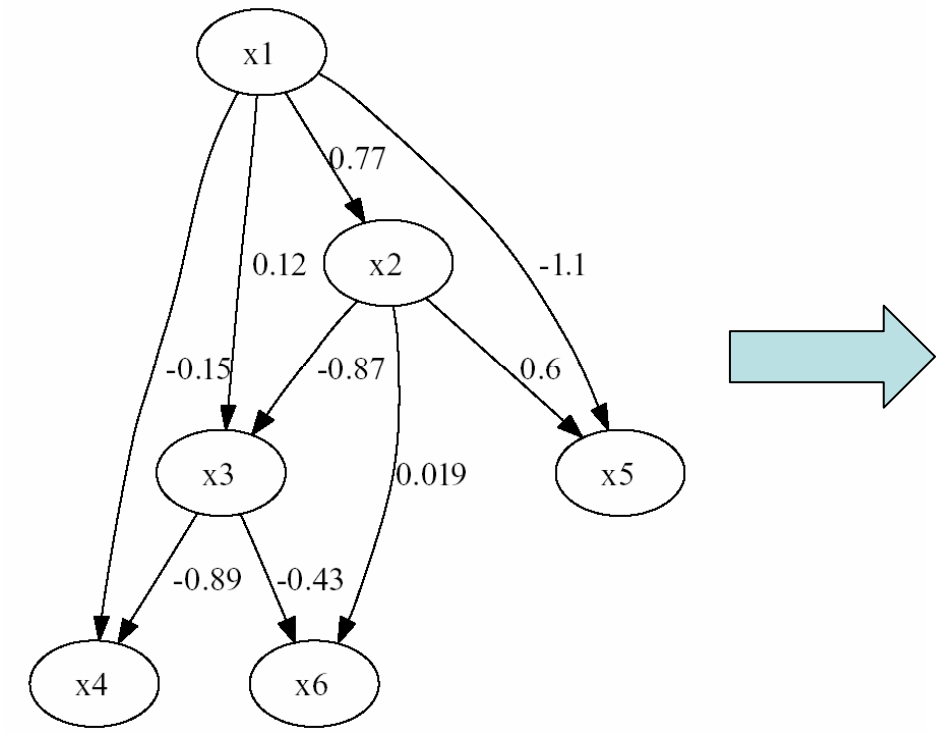
$$x_i = \sum_{k(i)>k(j)} b_{ij} x_j + e_i \quad \text{or} \quad \mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

Our goal

- We know
  - Data X is generated by $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$
- We do NOT know
  - Connection strengths: B
  - Order: k(i)
  - Disturbances: ei
- What we observe is data X only

- **Goal**
  - Estimate B and k using data X only!

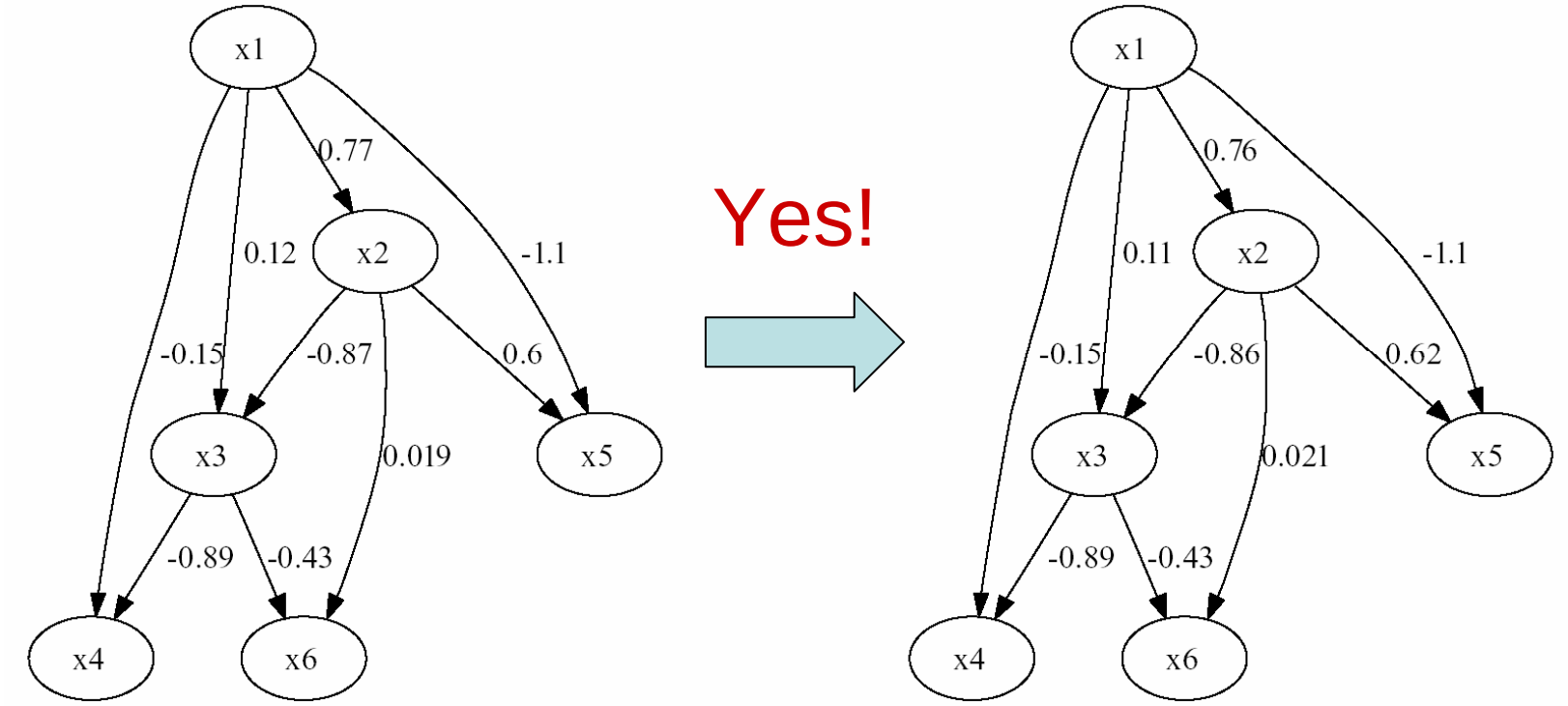
Can we recover the original network?



Original network

?

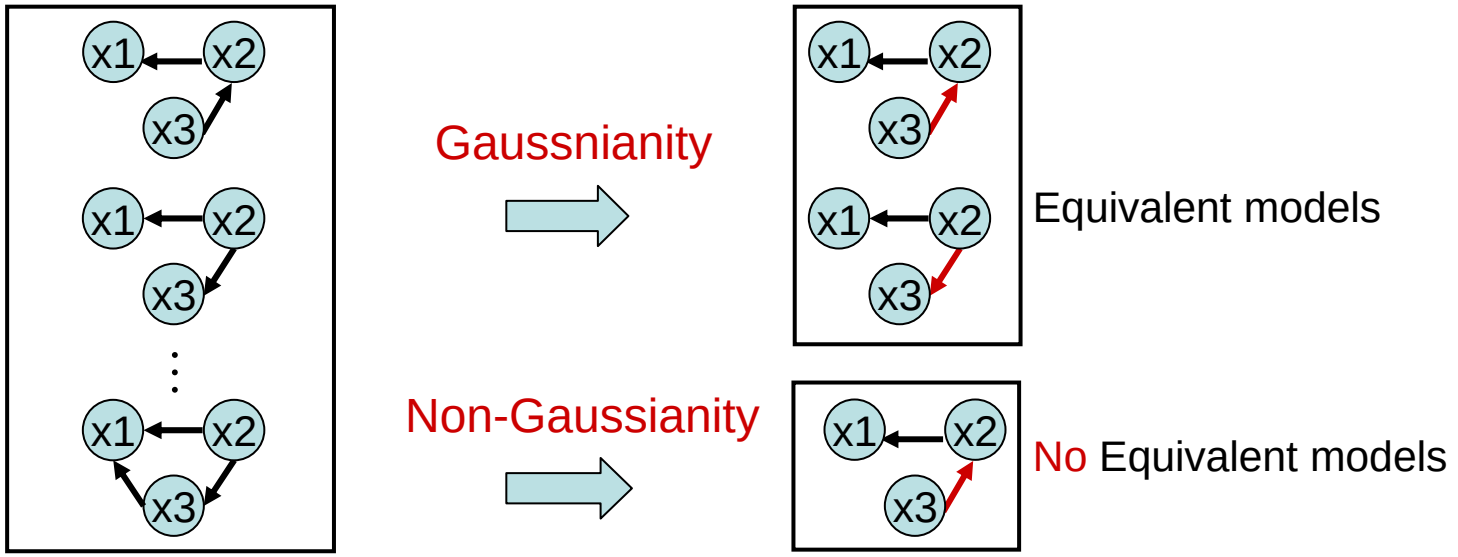
Can we recover the original network using ICA?



Original network **Estimated network**

Discovery of linear acyclic models from non-experimental data

- Existing methods (Bollen,1989; Pearl, 2000; Spirtes et al., 2000)
  - Gaussian assumption on disturbances ei
  - Produce many equivalent models
- Our LiNGAM approach (Shimizu et al, UAI2005, 2006 JMLR2)
  - Replace Gaussian assumption by non-Gaussian assumption
  - Can identify the connection strengths and structure



A simple example

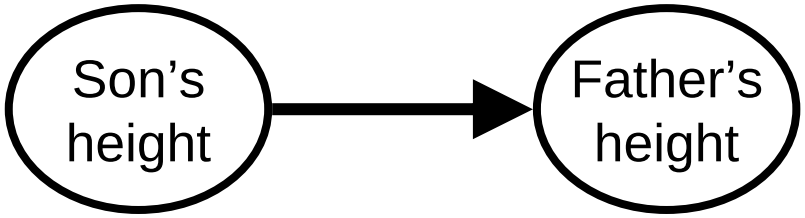
- Two regression models with opposite directions of arrows

- Linear-Gaussianity
  - Provide the same covariance structure
  - No way to distinguish between the two models

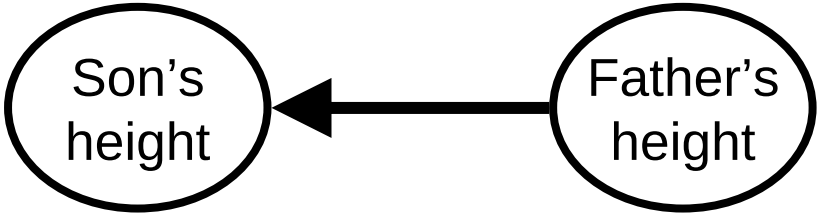
- Linear-Non-Gaussianity
  - Can evaluate model fit in terms of second- and fourth-order moments (Shimizu & Kano, 2006, JSPI)

---

p value = 0.04



p value = 0.53



126 samples

Linear Non-Gaussian Acyclic Models

(Shimizu et al, 2006 JMLR)

- As usual, linear acyclic models,
  but disturbances e_i are assumed to be non-gaussian:

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

- Note that B can be permuted to be strictly lower triangular if order k is known due to the DAG assumption (Bollen, 1989)
- No `faithfulness' or `stability'

Basic insight

- Observed variables x_i are linear combinations of non-gaussian independent disturbances e_i

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

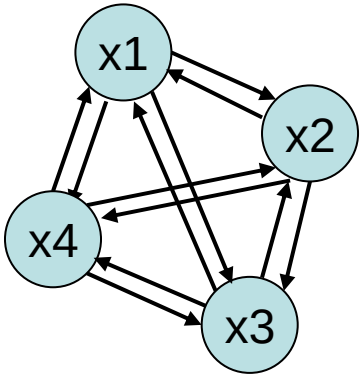
$$\Rightarrow \quad \mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}$$

$$= \mathbf{A}\mathbf{e}$$

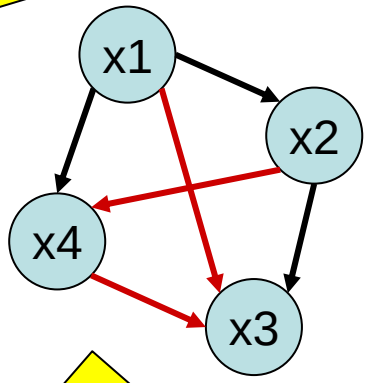
- Hence, we have a classic case of ICA

Outline of LiNGAM algorithm

1. Estimate B by ICA
+ post-processing

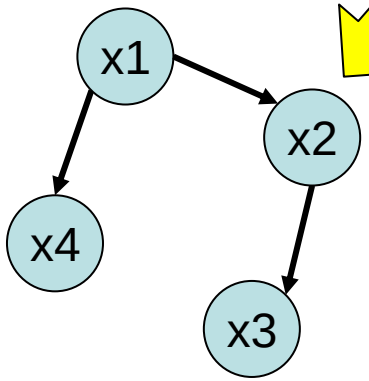


2. Find an order k(i) (DAG)



3. Prune redundant edges

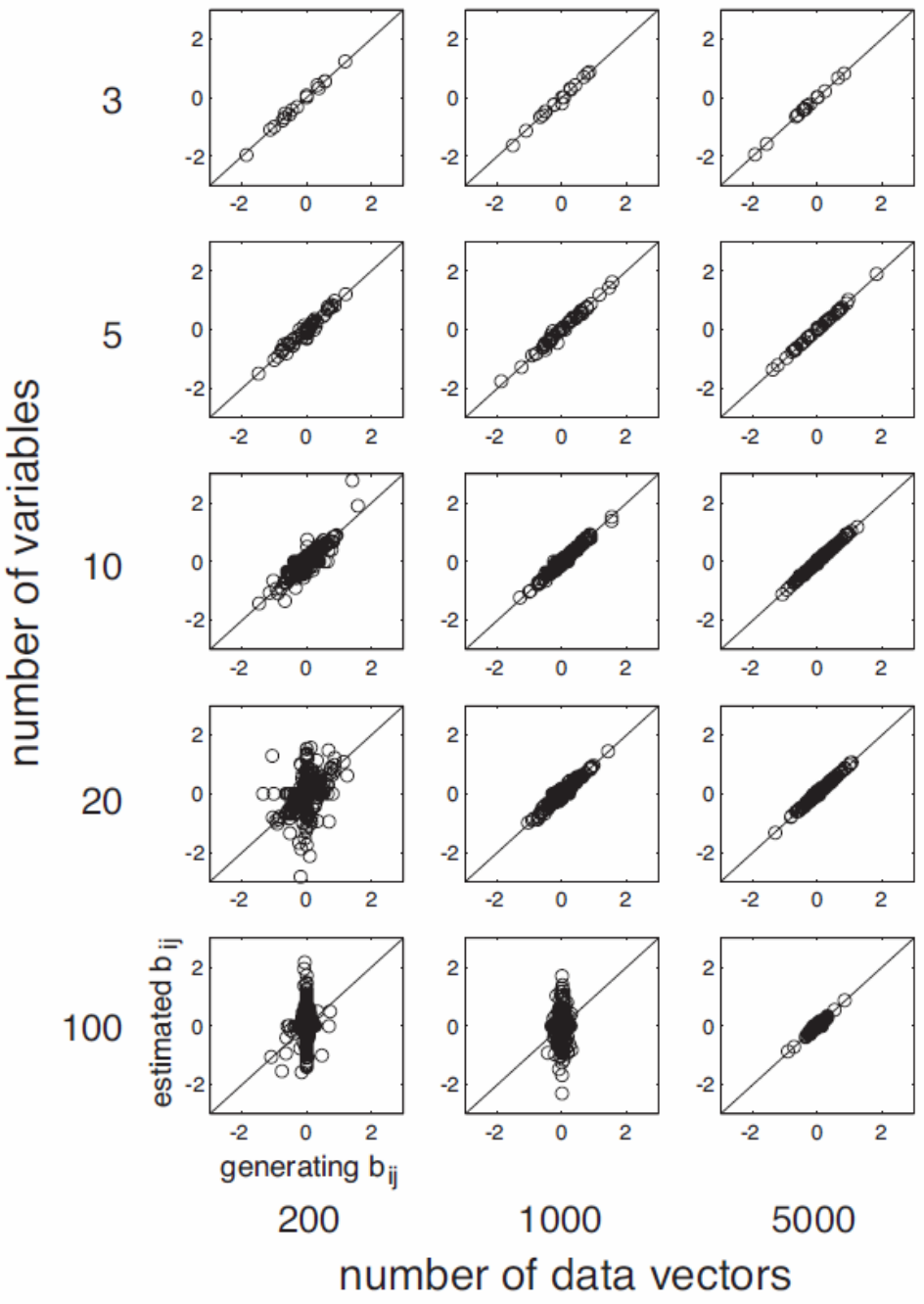
Wald test, Bootstrapping+OLS, etc.



Simulations

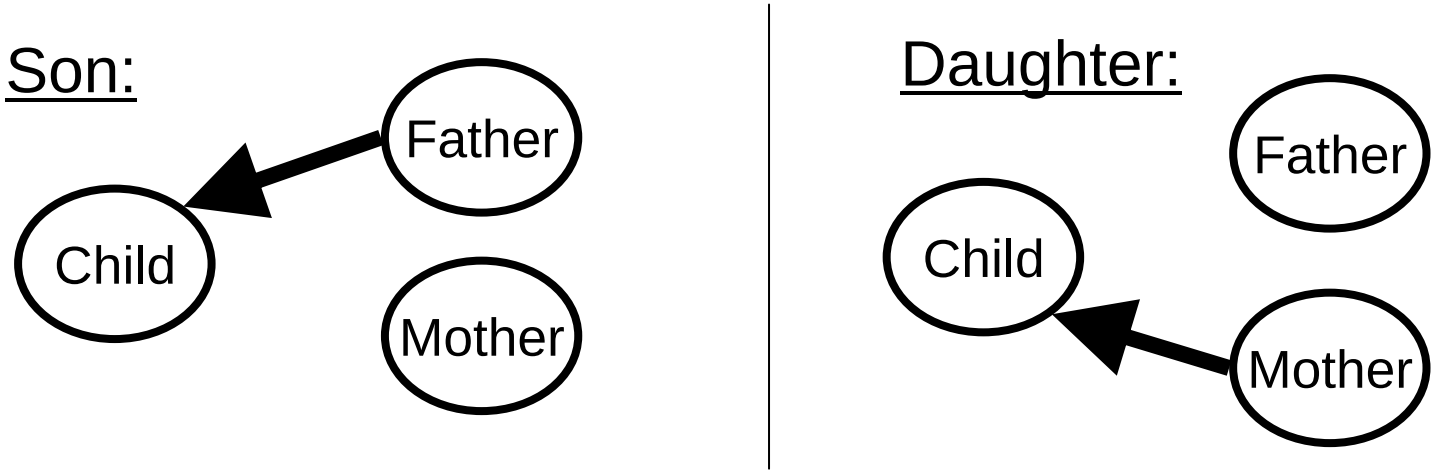
- Randomly generated B
- Various non-gaussian disturbances

- More along the main diagonals
  → more accurate estimation



An extension to cases where latent classes are present

- Hereditary effect on height (Kimura, 1974)



- Different network structures between different classes
- Often difficult to find class-membership
- Need a method to estimate in a data-driven way
  - latent classes of samples that are similar
  - the number of classes

Model estimation by ICA mixtures

- LiNGAM model for each class q

$$\mathbf{x} = \mathbf{B}_q \mathbf{x} + (\mathbf{I} - \mathbf{B}_q)\boldsymbol{\mu}_q + \mathbf{e}_q$$

Mean vector of x

- Estimable by ICA mixtures (Lee et al, 2000)
  - Min. Beta-divergence method (Mollah et al, 2006) learns A_q as well as the number of latent classes

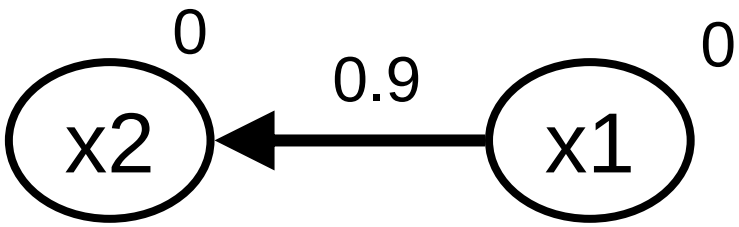
$$p(\mathbf{x} \mid \Theta) = \sum_{q=1}^{Q} p_q(\mathbf{x} \mid \boldsymbol{\mu}_q, \mathbf{A}_q) p(C = q),$$

$$\textbf{where } \mathbf{A}_q = (\mathbf{I} - \mathbf{B}_q)^{-1}$$

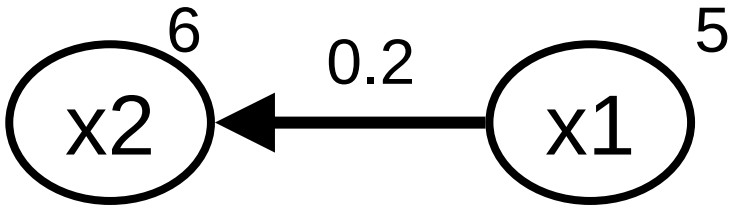
Illustrative example

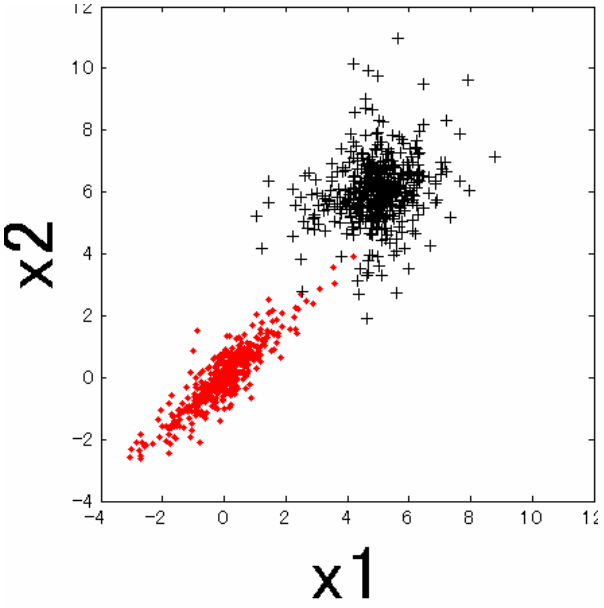
- Two classes have
  - the same causal order x1→x2
  - Different connection strengths between the classes (interaction effects)
  - Different means of x1 and x2
- Laplace distribution

Class 1:



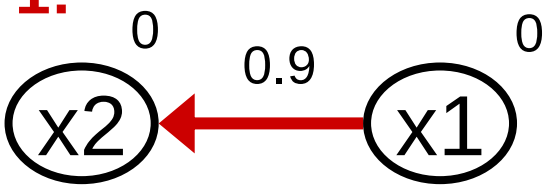
Class 2:





True:

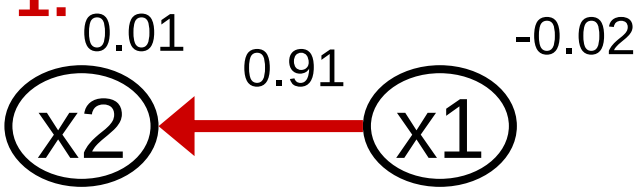
Class 1:



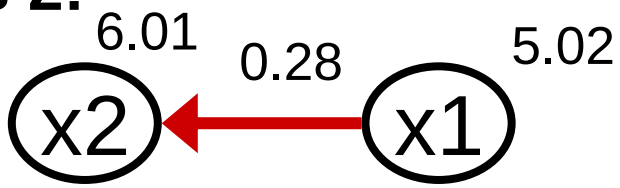
Class 2:

Latent class LiNGAM:

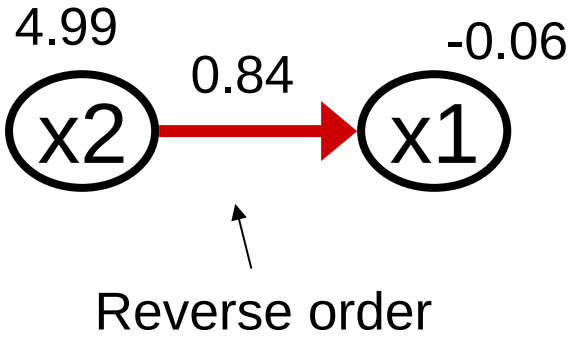
Class 1:



Class 2:



Regular LiNGAM:



Conclusions

- Discovery of linear acyclic models from non-experimental data is an important topic of current research
- A common assumption is linear-Gaussianity, but this leads to a number of indistinguishable models
- A non-Gaussian assumption allows all the connection strengths and structure of linear acyclic models to be identified (LiNGAM algorithm)
- In this contribution, we extended the method to the cases where latent classes are present (a nonlinear extension)
- Matlab/Octave code of the regular LiNGAM: http://www.cs.helsinki.fi/group/neuroinf/lingam/