

I. Model learning III

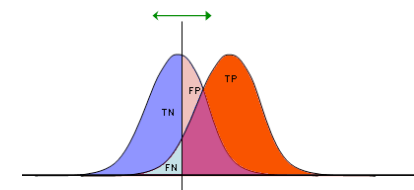
Today

- Behavior of independence tests with respect to amount of data
- Examples of learning in Tetrad
- Learning linear non-Gaussian causal models
- Examples of learning non-Gaussian networks

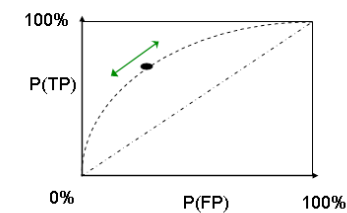
Behavior of independence tests with respect to amount of data

- Goal is to correctly classify independence (I) vs dependence (D)
- Four possible situations: $I \rightarrow I$, $D \rightarrow D$, $I \rightarrow D$, and $D \rightarrow I$ (symbol on left of arrow denotes true value, symbol on right of arrow the result of our test)
- The significance level (e.g. $\alpha = 5\%$), which we specify, equals (by definition) the probability that a true independence will be misclassified as dependent ('false positive')
- It also indirectly contributes to the value of β , the probability that a true dependence is classified as independent ('false negative'). The smaller we choose α the larger we will get β
- The relationship between α and β can be plotted in a ROC (receiver operating characteristic) curve

see: http://en.wikipedia.org/wiki/Receiver_operating_characteristic



TP	FP
FN	TN
1	1

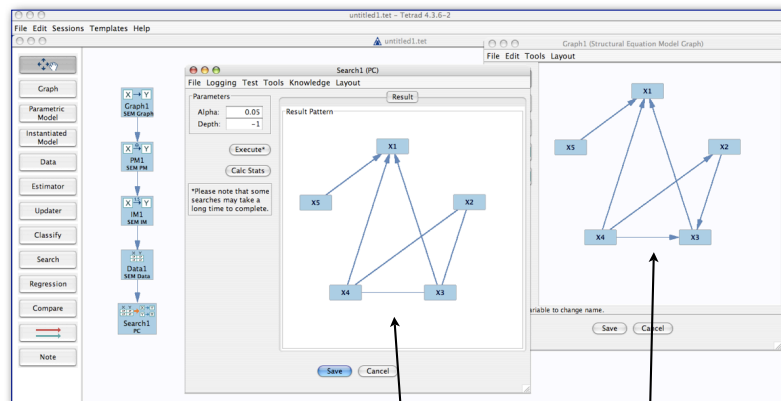


- If we keep the same 'false positive' rate α then as the amount of data grows we still get (by definition) the same number of false positives, but the probability β of 'false negatives' decreases towards zero
- Therefore, if we want to converge towards 'no classification errors' when the amount of data grows we should decrease α appropriately so that both α and β converge towards zero

Examples of model learning in Tetrad IV

- Freely available software by CMU group:
http://www.phil.cmu.edu/projects/tetrad_download/
- **Live test** of this software at the lecture...
 - discrete and continuous-gaussian
 - amount of data required
 - effect of p-value (significance level in independence tests)
 - effect of hidden variables

[screenshot of Tetrad IV]



learned pattern

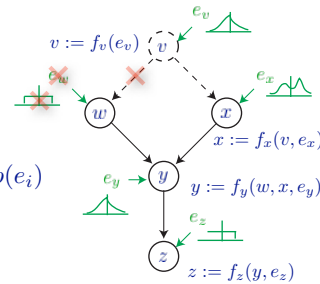
original DAG

Learning linear non-Gaussian causal models

Reminder:

Functional causal model

- Set of variables \mathbf{V}
- Observed variables $\mathbf{O} \subset \mathbf{V}$
- DAG over \mathbf{V}
- Mutually independent 'error' or 'disturbance' variables \mathbf{E} , distributions $p(e_i)$
- Deterministic functional relationships (autonomous mechanisms) f_i
- Key point: **interventions correspond to replacing single mechanisms**, i.e. only local effects (e.g. setting $w := w_0$ removes f_w but changes nothing else in the model)



Linear causal models (SEM)

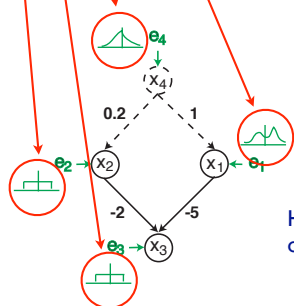
- Remember that in a **linear** causal model all mechanisms (functions) are linear. That is, we have

$$x_i := \sum_{j \in \text{pa}(i)} b_{ij}x_j + e_i$$

- Now, in a **linear non-gaussian** causal model additionally the error variables e_i have **non-normal distributions**
- We can never directly observe the e_i
- Some of the x_i can be latent (hidden) variables

Note: non-Gaussian!

Example



$$\begin{aligned} x_4 &:= e_4 \\ x_2 &:= 0.2x_4 + e_2 \\ x_1 &:= x_4 + e_1 \\ x_3 &:= -2x_2 - 5x_1 + e_3 \end{aligned}$$

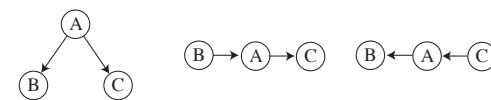
Here, x_4 is a hidden variable, and we only observe $\{x_1, x_2, x_3\}$

data: (i.i.d. sample vectors)

x_1																				
x_2																				
x_3																				

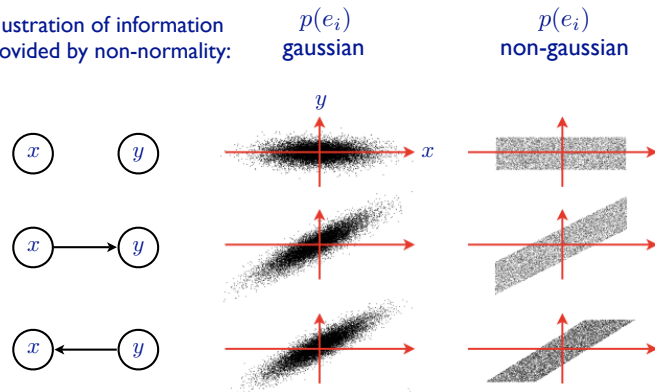
Estimation based on conditional independencies

- So far in the course we have only considered learning methods purely based on conditional independencies
- Such methods lead to the problem of d-separation-equivalence
- When disturbances are Gaussian then there is nothing more we can do. **When disturbances are non-Gaussian one can still (of course) use these methods but the results are then limited to d-separation-equivalence.**
- Example: Three independence-equivalent models...



Estimation based on non-Gaussianity

Illustration of information provided by non-normality:



Estimation using ICA: no hidden variables case

(Shimizu et al, UAI 2005; Shimizu et al, JMLR 2006)

- Assume that there are no hidden variables. We have...

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

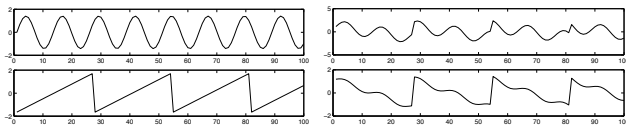
where the e_i are non-gaussian and independent, by the assumptions of the model. Hence we have a classic case of **Independent Component Analysis (ICA)** with a square invertible mixing matrix \mathbf{A} . We can estimate the mixing matrix from the data!

Independent Component Analysis, ICA

- Roots: Blind source separation

$$x_1(t) = a_{11}s_1 + a_{12}s_2$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2$$



Cocktail-party demo:



- The mixing process can be written in the form $\mathbf{X} = \mathbf{A}\mathbf{S}$
- Can we, based on the observed data \mathbf{X} , estimate the mixing matrix \mathbf{A} and the original sources \mathbf{S} ?

Answer: Yes! (if the signals are independent, non-Gaussian, and the mixing matrix is invertible)

The answer is obtained by looking for a matrix \mathbf{W} such that the rows of $\hat{\mathbf{S}} = \mathbf{W}\mathbf{X}$ are as independent as possible (Comon, 1994; Jutten and Héroult, 1991; Hyvärinen et al, 2001)

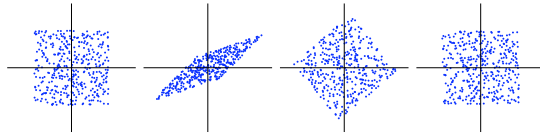
But... permutation and scaling indeterminacy (see later)

- Basic idea: Searching for non-gaussian projections!

Illustration

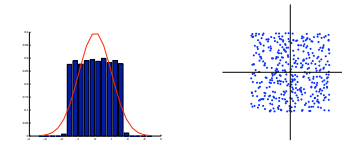
Two components with uniform distributions:

Original components, observed mixtures, PCA, ICA

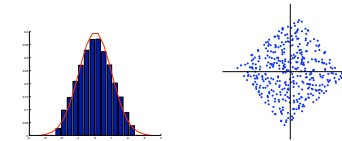


PCA does not find original coordinates, ICA does!

Illustration of changes in nongaussianity



Histogram and scatterplot, original uniform distributions



Histogram and scatterplot, mixtures given by PCA

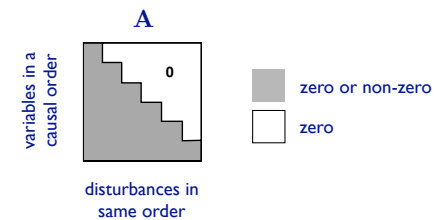
The scaling and permutation indeterminacy in ICA:

$$\mathbf{X} = \mathbf{A}\mathbf{S}$$

- Multiplying a row of \mathbf{S} with the factor k and dividing the corresponding column of \mathbf{A} by the same factor k does not change the observed data at all. So, it is impossible to uniquely recover the scale of the independent components. Usually they are assumed unit variance
- Permuting the rows of \mathbf{S} and correspondingly permuting the columns of \mathbf{A} also does not change the data at all. So it is impossible to recover the order of the independent components. We will obtain some random permutation of the original order

The permutation problem has to be solved when learning linear models (we must pair up the variables and their disturbances!)

- Note: in ICA, the columns of \mathbf{A} have no defined order, and the rows are in the same order as the data. But if the model holds it can be permuted to lower-triangular form!



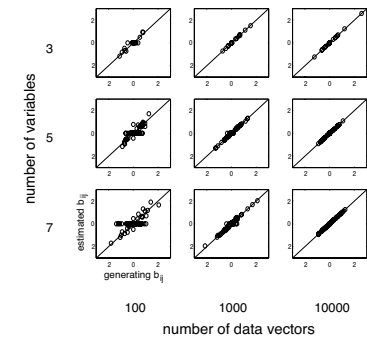
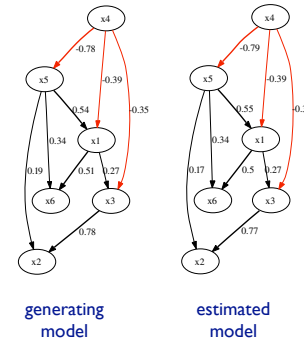
- If ICA yielded exact estimates then finding the appropriate permutations (ordering of rows and columns of the mixing matrix) would be trivial
- Since the estimates are just estimates, we do not know which elements should be zero and which non-zero, hence we need to optimize some measure of triangularity over the permutations
- For large numbers of variables, brute-force impossible. However, we can find fast methods which work quite well

Possible, in principle and in practice, to estimate the complete linear causal model, including all parameters!



Full Matlab package available:

<http://www.cs.helsinki.fi/group/neuroinf/lingam/>



Hidden variables → overcomplete ICA basis!

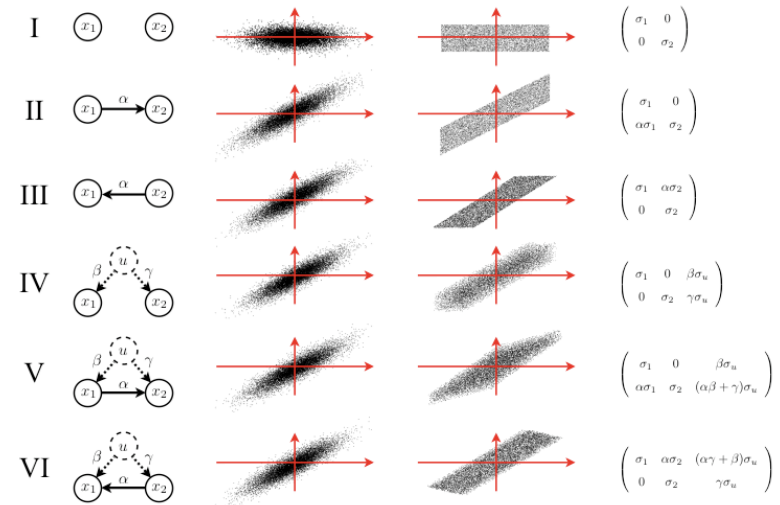
- 'One or more rows are missing' in the data matrix...

data: (i.i.d. sample vectors)

x_1																				
x_2																				
x_3																				
x_4																				

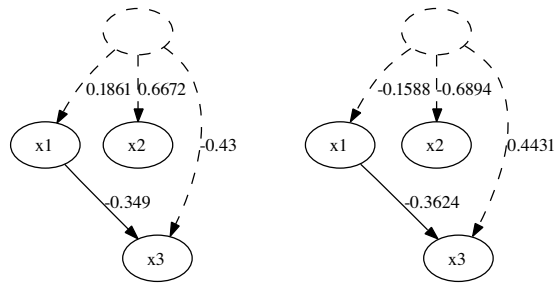
mixing matrix A :

- We have a case of overcomplete ICA: 'more sources than sensors'. The mixing matrix A has more columns than rows
- Overcomplete ICA is identifiable (Eriksson & Koivunen, 2004) but in practice much more difficult than 'regular' ICA



Experiments

Example:



Summary of linear, non-Gaussian causal discovery

- Linear causal models (SEMs) are in widespread use
- Estimation methods based on conditional independencies are restricted to what is possible for Gaussian distributed data
- But: if the data is non-Gaussian, then full identification is achieved if no unobserved confounders exist
- When there are hidden variables...
 - The model is identifiable up to a finite set of observationally equivalent models
 - Practical implementation much more difficult than for the no-hidden-variables case. However, initial results show that the method is feasible

Required reading:

- Shimizu et al (2005):
“Discovery of non-gaussian linear causal models using ICA”
<http://www.cs.helsinki.fi/u/phoyer/papers/pdf/uai2005cameraready.pdf>

Links:

- Tetrad IV:
http://www.phil.cmu.edu/projects/tetrad_download/
- Brief description and demo of ICA:
<http://www.cis.hut.fi/projects/ica/icademo/>
http://www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi
- LiNGAM homepage:
<http://www.cs.helsinki.fi/group/neuroinf/lingam/>