

Causal discovery with non-Gaussian noise: LiNGAM and OT-LiNGAM

Lecture 6 – Causality and Graphical Models

Christophe Ambroise

M2 Data Science – Université d'Évry

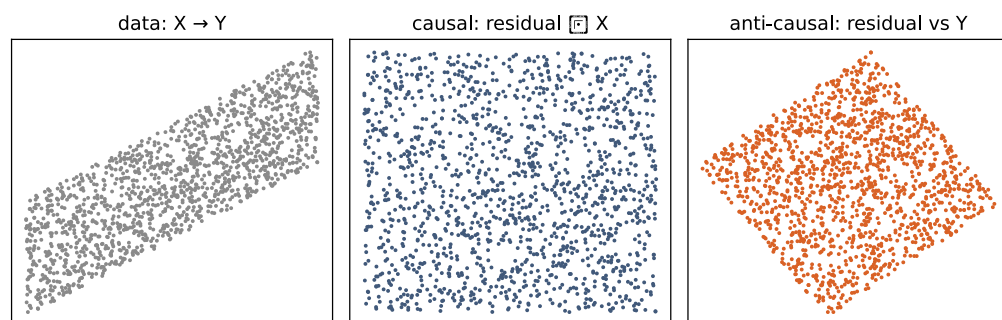
Lectures 4 and 5

- SCM $V_j := f_j(V_{pa(j)}, U_j)$; intervention = graph surgery
- Backdoor adjustment; instrument and Wald ratio
- Structure learning: constraint-based (PC), score-based (BIC, GES)
- PC output: CPDAG, Markov equivalence class
- Gaussian linear SEM: direction of $X_1 - X_2$ not identifiable
- SHD: additions + deletions + reversals

Prior knowledge

Beyond Markov equivalence:
LiNGAM

Two variables, one direction?



- Both residuals **uncorrelated** with the regressor
- Only the causal one **independent**: non-Gaussianity breaks the symmetry

Darmois–Skitovich theorem

If two linear combinations of independent variables are independent, every variable appearing in both is Gaussian.

$$\varepsilon_j \text{ independent, } \sum_j a_j \varepsilon_j \perp \sum_j b_j \varepsilon_j \implies \forall j : a_j b_j \neq 0 \Rightarrow \varepsilon_j \text{ Gaussian}$$

- Anti-causal residual $X - \hat{b}Y$ and Y : both mix ε_X
- Non-Gaussian $\varepsilon_X \rightarrow$ dependence $\rightarrow$ wrong direction detected

Linear non-Gaussian acyclic model

In LiNGAM, the data follow a linear SEM on a DAG, with mutually independent noises of which at most one is Gaussian.

$$X = BX + \varepsilon, \quad B \text{ acyclic (strictly lower triangular up to permutation),}$$

Identifiability (Shimizu *et al.*, 2006)

Under these assumptions, B and the causal order are identifiable from the distribution of X alone.

Link with independent component analysis

- $X = (I - B)^{-1}\varepsilon = A\varepsilon$: linear mixture of independent sources
- ICA: unmixing $W = I - B$ recovered up to **scaling and permutation**
- Acyclicity fixes the permutation: diagonal of $I - B$ without zero

ICA-LiNGAM

1. FastICA $\rightarrow \hat{W}$
2. Row permutation: no zero on the diagonal (Hungarian algorithm)
3. Rows normalised to unit diagonal $\rightarrow \hat{B} = I - \hat{W}$
4. Permutation making $\hat{B}$ closest to lower triangular $\rightarrow$ causal order

DirectLiNGAM

X_j is exogenous (a source) if and only if every other variable, regressed on X_j , leaves a residual independent of X_j .

$$X_j \text{ source} \iff \forall i \neq j: X_i - \frac{\text{Cov}(X_i, X_j)}{\text{Var } X_j} X_j \perp X_j$$

1. Most independent candidate $\rightarrow$ next variable of the order
2. Its effect regressed out of all remaining variables
3. Repeat; then B by adaptive lasso along the order

Shimizu et al. (2011), *JMLR*. Python package [lingam](#).

DirectLiNGAM in practice

```
1 import numpy as np, lingam
2 rng = np.random.default_rng(1); n = 2000
3 e = rng.laplace(size=(n, 4)) # non-Gaussian noises
4 x0 = e[:, 0]
5 x2 = 0.8 * x0 + e[:, 2]
6 x1 = -0.7 * x2 + e[:, 1]
7 x3 = 0.5 * x0 + 0.6 * x1 + e[:, 3]
8 X = np.c_[x0, x1, x2, x3]
9 model = lingam.DirectLiNGAM().fit(X)
10 print("order:", [int(i) for i in model.causal_order_])
11 print(np.round(model.adjacency_matrix_, 2)) # B[j, k] : effect k -> j

order: [0, 2, 1, 3]
[[ 0.  0.  0.  0. ]
 [ 0.  0. -0.69 0. ]
 [ 0.79 0.  0.  0. ]
 [ 0.51 0.6  0.  0. ]]
```

OT-LiNGAM: non-Gaussianity by optimal transport

Measuring non-Gaussianity

- Classical contrasts: kurtosis, negentropy approximations, likelihood surrogates
- Choice of a nonlinearity or moment order $\rightarrow$ method tuned to source shape
- Wanted: **one** criterion, no tuning, only finite second moments

Laplante, Ambroise & Humbert, *Contrast-free ICA and causal inference via Wasserstein distances to the Gaussian*, preprint. Package [otlingam](#).

2-Wasserstein distance

The squared 2-Wasserstein distance is the smallest mean squared displacement among all couplings of the two distributions.

$$\mathcal{W}_2(\mu, \nu)^2 = \inf_{X \sim \mu, Y \sim \nu} \mathbb{E}[(X - Y)^2] \stackrel{\text{dim } 1}{=} \int_0^1 (F_\mu^{-1}(u) - F_\nu^{-1}(u))^2 du$$

- Dimension 1: optimal coupling = matching of quantiles
- Non-Gaussianity of X : $\mathcal{W}_2(\text{std}(X), \mathcal{N}(0, 1))^2$, zero iff X Gaussian

Exact empirical computation

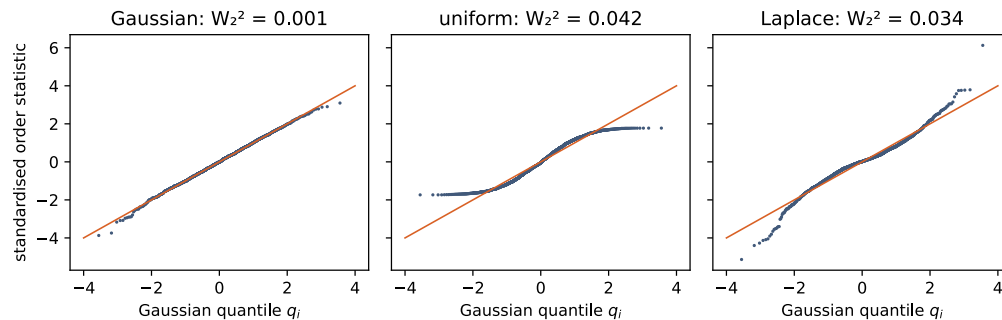
For a sample $x^{(1:n)} \leq \dots \leq x^{(n:n)}$, the distance to $\mathcal{N}(0, 1)$ is a sum of squares between order statistics and Gaussian quantile averages.

$$\mathcal{W}_2(\mu_n, \mathcal{N}(0, 1))^2 = \frac{1}{n} \sum_{i=1}^n (x^{(i:n)} - q_i)^2 + c_n, \quad q_i = n(\phi(z_{i-1}) - \phi(z_i)),$$

- $c_n = 1 - \frac{1}{n} \sum_i q_i^2$: constant
- Cost $O(n \log n)$: one sort

13

Non-Gaussianity of three distributions



Mixtures are closer to the Gaussian

Under independence and standardisation, any non-trivial normalised mixture of sources is strictly closer to the Gaussian than the corresponding weighted average of the sources' distances.

$$\|\alpha\|_2 = 1, \#\{j : \alpha_j \neq 0\} > 1 \implies \mathcal{W}_2\left(\sum_j \alpha_j S_j, \mathcal{N}(0, 1)\right)^2 < \sum_j \alpha_j^2 \mathcal{W}_2(S_j, \mathcal{N}(0, 1))^2$$

Assumptions: S_j independent, centred, unit variance, at most one Gaussian.

15

14

16

Wasserstein order objective

For a candidate order σ , regress each variable on its predecessors and score the non-Gaussianity of the standardised residuals.

$$G(\sigma) = \sum_{j=1}^d \mathcal{W}_2\left(\text{std}(R_j(\sigma)), \mathcal{N}(0, 1)\right)^2, \quad R_j(\sigma) = X_j - \text{proj}_{\text{span}\{X_k: k \text{ before } j \text{ in } \sigma\}}$$

- Causal order: residuals = structural noises ε_j
- Wrong order: some residual mixes several noises $\rightarrow$ closer to Gaussian

Consistency

- Empirical objective $\hat{G}_n$: OLS residuals + exact $\mathcal{W}_2$
- Uniform convergence: $\max_{\sigma} |\hat{G}_n(\sigma) - G(\sigma)| \rightarrow 0$ a.s., finite second moments only
- Margin $\Gamma = \max G - \max_{\sigma \notin \mathcal{I}^*} G > 0$

$$\mathbb{P}(\hat{\sigma}_n \notin \mathcal{I}^*) \leq \mathbb{P}\left(\max_{\sigma} |\hat{G}_n(\sigma) - G(\sigma)| \geq \Gamma/2\right) \rightarrow 0$$

- Sub-Gaussian data: rate $d^{5/4}n^{-1/4}$ up to log factors

Identification of causal orders

Theorem (Laplante, Ambroise & Humbert)

Under independent, non-degenerate noises with at most one Gaussian, the maximisers of G are exactly the causal orders of the model.

$$\arg \max_{\sigma \in \mathfrak{S}_d} G(\sigma) = \mathcal{I}^* = \{\sigma : \Pi^{\sigma} B (\Pi^{\sigma})^{\top} \text{ strictly lower triangular}\}$$

Key steps: $\text{std}(R(\sigma)) = Q^{\sigma} \text{std}(\varepsilon)$ with Q^{σ} orthogonal; strict subadditivity; Q^{σ} signed permutation $\Rightarrow \sigma$ causal.

Exhaustive solver: dynamic programming

Decomposable score $\rightarrow$ recursion on subsets U , $H(U)$ = best score of an order of U :

$$H(U) = \max_{u \in U} \left\{ H(U \setminus \{u\}) + \hat{g}_u(U \setminus \{u\}) \right\}, \quad H(\emptyset) = 0$$

- $\hat{g}_u(P)$: non-Gaussianity of the residual of X_u on X_P
- Backtracking of the stored sinks $\rightarrow$ order
- $d 2^{d-1}$ local scores instead of $d!$ orders; exact optimum up to $d \approx 30$

Greedy solver and full pipeline

Greedy OT-LiNGAM: most non-Gaussian standardised residual $\rightarrow$ next source; its effect regressed out; repeat. Cost $O(d^2 n \log n)$.

Solver	Cost	Global optimum
Exhaustive OT-LiNGAM	$O(d^2 d^d (nd^2 + n \log n))$	yes
Greedy OT-LiNGAM	$O(d^2 n \log n)$	no
OT-ICA-LiNGAM	$O(nd^2 + dn \log n)$ / iteration	no

Order known $\rightarrow B$ by adaptive lasso along the order.

Evaluation: disorder of an order

Number of true edges reversed by the estimated order:

$$\text{dis}(\hat{\sigma}) = \#\{(k, j) : B_{jk} \neq 0, \hat{\sigma}^{-1}(k) > \hat{\sigma}^{-1}(j)\}$$

- $\text{dis}(\hat{\sigma}) = 0 \iff \hat{\sigma}$ causal order
- Simulations of the paper (ER and scale-free graphs, Laplace or Student noises): exhaustive OT-LiNGAM close to zero disorder, robust when some noises approach Gaussianity
- Continuous methods (DAGMA): degraded under heteroscedastic noises

OT-LiNGAM in practice

```

1 import numpy as np
2 from otlingam import ExhaustiveOTLiNGAM, GreedyOTLiNGAM
3 from otlingam.utils import disorder
4 rng = np.random.default_rng(1); n = 2000
5 e = rng.laplace(size=(n, 4))
6 x0 = e[:, 0]; x2 = 0.8 * x0 + e[:, 2]
7 x1 = -0.7 * x2 + e[:, 1]; x3 = 0.5 * x0 + 0.6 * x1 + e[:, 3]
8 X = np.c_[x0, x1, x2, x3]
9 B_true = np.zeros((4, 4)); B_true[2, 0], B_true[1, 2], B_true[3, 0], B_true[3, 1] = .8, -.7, .5, .6
10 for M in (ExhaustiveOTLiNGAM(), GreedyOTLiNGAM()):
11     M.fit(X)
12     print(type(M).__name__, [int(i) for i in M.causal_order_], "disorder:", disorder(B_true, M.causal_order_))

```

ExhaustiveOTLiNGAM [0, 2, 1, 3] disorder: 0
 GreedyOTLiNGAM [0, 2, 1, 3] disorder: 0

LiNGAM analysis in practice

Non-Gaussianity check

```
1 from scipy import stats
2 print("excess kurtosis:", np.round(stats.kurtosis(X), 2))
3 print("Shapiro p-values:", [f"{stats.shapiro(c).pvalue:.1e}" for c in X.T])
```

```
excess kurtosis: [2.72 0.99 1.57 1.21]
Shapiro p-values: ['4.8e-21', '5.6e-09', '2.0e-15', '2.5e-10']
```

- Kurtosis far from 0, tiny Shapiro p-values: non-Gaussian variables
- Near-Gaussian variables: directions unreliable
- Discrete or binary variables: LiNGAM only an approximation

Bootstrap stability of edges

```
1 model = lingam.DirectLiNGAM().fit(X)
2 boot = lingam.DirectLiNGAM().bootstrap(X, n_sampling=100) # separate instance: refitted in place
3 P = boot.get_probabilities(min_causal_effect=0.05) # P[j, k]: frequency of edge k -> j
4 print(np.round(P, 2))
```

```
[[0.  0.  0.  0. ]
 [0.01 0.  1.  0. ]
 [1.  0.  0.  0. ]
 [1.  1.  0.02 0. ]]
```

- One fit = one estimate; bootstrap = frequency of each edge over resamples
- Edges reported **with** their frequency

25

26

Independence of residuals

```
1 print(np.round(model.get_error_independence_p_values(X), 2))
2 rng = np.random.default_rng(1); h = rng.laplace(size=n) # h: hidden common cause
3 a = h + rng.laplace(size=n); b = 0.5 * a + h + rng.laplace(size=n)
4 D = np.c_[a, b]; m = lingam.DirectLiNGAM().fit(D)
5 pv = m.get_error_independence_p_values(D)[0, 1]
6 print(f"a -> b: {m.adjacency_matrix_[1, 0]:.2f}, p-value {pv:.0e}")
```

```
[[0.  0.82 0.12 0.17]
 [0.82 0.  0.35 0.74]
 [0.12 0.35 0.  0.67]
 [0.17 0.74 0.67 0.  ]]
a -> b: 0.97, p-value 3e-06
```

- Model assumption: independent noises → residuals pairwise independent
- Small p-value: hidden confounder, non-linearity or feedback
- Hidden *h*: effect **0.97** instead of **0.5**, residuals clearly dependent

Prior knowledge

```
1 pk = lingam.utils.make_prior_knowledge(4, exogenous_variables=[0], sink_variables=[3])
2 print(pk) # pk[j, k]: 1 path k -> j, 0 no path, -1 unknown
3 print([int(i) for i in lingam.DirectLiNGAM(prior_knowledge=pk).fit(X).causal_order_])
```

```
[[ -1  0  0  0]
 [ -1 -1 -1  0]
 [ -1 -1 -1  0]
 [ -1 -1 -1 -1]]
[0, 2, 1, 3]
```

- Exogenous variables: genotype, treatment, time
- Sinks: final phenotype, outcome
- Fewer orders to search, fewer errors

27

28

Total causal effect

The total effect of X_k on X_j sums the contributions of all directed paths from X_k to X_j : the change of X_j under an intervention $do(X_k = x + 1)$ instead of $do(X_k = x)$.

```
1 print(f"x0 -> x3: {model.estimate_total_effect(X, 0, 3):.2f}")
```

x0 -> x3: 0.18

- True value: direct 0.5 + path $x_0 \rightarrow x_2 \rightarrow x_1 \rightarrow x_3$: $0.8 \times (-0.7) \times 0.6 \rightarrow 0.164$
- Bootstrap version: `boot.get_total_causal_effects()`

Case study: chloroplast RNA maturation

Estimators and report

Estimator	Package	Use
DirectLiNGAM	lingam	bootstrap, prior knowledge, tests
ExhaustiveOTLiNGAM	otlingam	exact search, few variables
GreedyOTLiNGAM	otlingam	many variables

Report of a LiNGAM analysis

Non-Gaussianity; orders from two estimators; bootstrap frequencies of the edges; residual tests; prior knowledge; total effects.

Biological context

- Chloroplast transcripts of *Arabidopsis thaliana*: genes *ndhB*, *ndhD*
- Post-transcriptional maturation: C-to-U **editing** at specific sites, intron **splicing**
- Nanopore long reads: one read = one full-length RNA molecule
- Each read: snapshot of one molecule at some stage of its maturation

Data: Guilcher et al. (2021), *Int. J. Mol. Sci.* 22, 11297.

Data: one binary vector per read

	ed1	ed2	ed3	ed4	ed5
	False	False	False	True	.
	True	False	.	.	.
	False	False	False	False	.
	False	True	False	True	.
	True	False	False	True	True
	False	True	True	True	.

Gene	Reads	Events	Complete reads
<i>ndhB</i>	1 899	12 editing sites + intron splicing	117
<i>ndhD</i>	7 574	5 editing sites	930

True: event observed; **False**: not observed; **Err**: sequencing error; empty: not covered.

Question: chronology of maturation events

- Order in which sites are edited and the intron spliced
- Chronology as a **causal order** of the event variables
- Snapshot data only: no time stamps

Open problem

Can a causal order learned from snapshots of molecules reveal the maturation chronology?

Methodological difficulties

- **Binary** variables: linear non-Gaussian model only an approximation
- **Missing blocks**: reads truncated, contiguous sites unobserved
- **Errors** (**Err**): measurement noise on some sites
- Rare events (sites edited in < 10 % of reads)
- Validation: no ground-truth chronology