Identifiability of Gaussian Structural Equation Models with Equal Error Variances

(J.Peters and P.Buhlmann, Biometrika, 2014)

Bohyeon Park

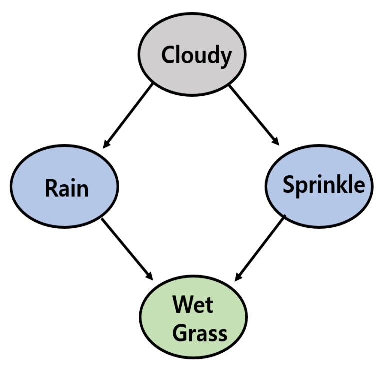
2019.01.10

Outline

Outline

1. Introduction

1. Introduction



1.1. Directed Acyclic Graphical(DAG) model

- Directed acyclic graph $\mathcal{G}$:

$$1 \longrightarrow 2 \longrightarrow 3$$

  - $\mathcal{G} = (V, E)$.
  - V : a set of nodes, e.g. $V = \{1, 2, 3\}$.
  - E : a set of directed edges, $E = \{(1, 2), (2, 3)\}$.

1.1. Directed Acyclic Graphical(DAG) model

- DAG model:

  - $X := (X_j)_{j \in V}$: a set of random variables, e.g. $X = \{X_1, X_2, X_3\}$.
  - DAG model has the factorization,

$$P(\mathcal{G}) = P(X_1, X_2, \ldots, X_p) = \prod_{j=1}^{p} P(X_j \mid X_{Pa(j)}).$$

  - In this graph,

$$P(\mathcal{G}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2).$$

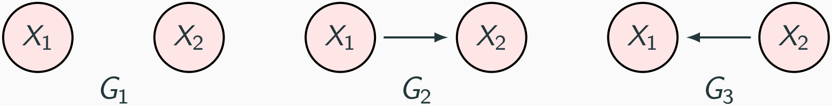
1.2. Identifiability from the distribution

- Problem:

Given the joint distribution $P(\mathcal{G})$, can we recover the graph $\mathcal{G}_0$?

"Negative!!"

- Reason:



  - We can distinguish G_2 and G_3 from G_1.
  - We cannot identify a direction of an edge. Hence, we cannot distinguish G_2 and G_3.

1.2. Identifiability from the distribution

- Exception case:
  1. Linear non-Gaussian SEMs (Shimizu et al., 2006).
  2. Non-parametric SEMs with additive independent noise (Peters et al.,2012).
  3. **Gaussian SEMs with the equal error variances (Peters and Bühlmann, 2013)**.

2. Identifiability for Gaussian SEM with equal error variances

2. Identifiability for Gaussian SEM with equal error variances

2. Identifiability for Gaussian SEM with equal error variances

- Gaussian SEM:

$$X_j = \Sigma_{k \in PA_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j \quad (j = 1, \ldots, p).$$

  - V: a set of nodes in a graph, $V = \{1, \ldots, p\}$.
  - X_j: random variable, $j \in V$.
  - N_j: noise term, $N_j \sim^{IID} N(0, \sigma^2)$ with $\sigma^2 > 0$.
  - $\beta_{jk} \neq 0$ for all $k \in PA_j^{\mathcal{G}_0}$, otherwise $\beta_{jk} = 0$.

2. Identifiability for Gaussian models with equal error variances

- Theorem 1.
  Let $P(\mathcal{G})$ be generated from model Gaussian SEM,
  Then $\mathcal{G}_0$ is identifiable from $P(\mathcal{G})$ and the coefficients β_{jk} can be reconstructed for all j and $k \in PA_j^{\mathcal{G}_0}$.

- Assumptions: **Non-zero coefficient**, **Causal sufficiency**.
  - Non-zero cofficient: $\beta_{jk} \neq 0$, $k \in PA_j^{\mathcal{G}_0}$.
  - Causal sufficiency: all variables are observed.

3. Penalized maximum likelihood estimator

3. Penalized maximum likelihood estimator

3. Penalized maximum likelihood estimator

- Penalized maximum likelihood estimator:

$$\{\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)\} = \arg\min_{\beta \in \mathcal{B}, \sigma^2 \in \mathbb{R}^+} -\ell\left(\beta, \sigma^2; X^{(1)}, \ldots, X^{(n)}\right) + \lambda ||\beta||_0, \quad (1)$$

where

$$-\ell\left(\beta, \sigma^2; X^{(1)}, \ldots, X^{(n)}\right) = \frac{np}{2}\log(2\pi\sigma^2) + \frac{n}{2\sigma^2}\mathrm{tr}\{(I-B)^T(I-B)\hat{\Sigma}\}.$$

  - B: p$\times$ p matrix with $B_{jk} = \beta_{jk}$.
  - $\hat{\Sigma}$: sample covariance matrix.
  - σ^2: error variance.
  - $\lambda = \log(n)/2$: the objective function in equation (1) is the BIC score.

3. Penalized maximum likelihood estimator

- Convergence rate:

  For $\lambda_n = \log(n)/2$, $n \to \infty$,

  $$\Sigma_{j,k=1}^{p}\{\hat{\beta}_{jk}(\lambda_n) - \beta_{jk}^{0}\}^2 = O_p\{\log(n)n^{-1}\}.$$

- Consistency:

  For $\lambda_n = \log(n)/2$, $n \to \infty$,

  $$\mathrm{pr}(\hat{\mathcal{G}}_n = \mathcal{G}_0) \to 1.$$

3. Penalized maximum likelihood estimator

- Computational complexity:

$$p = 20 \quad \rightarrow \quad 2 \cdot 3 \times 10^{72} \ \text{ DAG } \ \mathcal{G}.$$

$\Rightarrow$ **Greedy search algorithm**

4. Greedy search algorithm

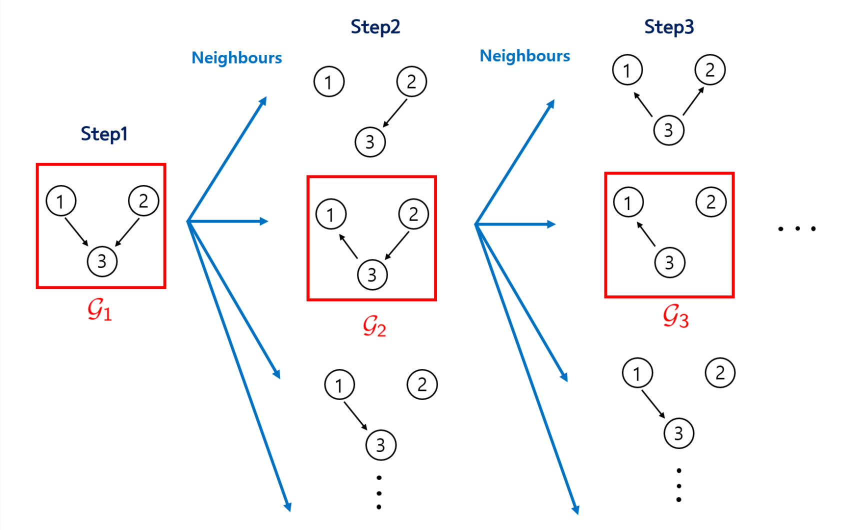
4. Greedy search algorithm

4. Greedy search algorithm

- We consider at least k neighbours.
- If the neighbouring $\mathcal{G}_t$ have a lower BIC score than $\mathcal{G}_{t-1}$, $\mathcal{G}_{t-1}$ move to $\mathcal{G}_t$.
- If all neighbours have a higher BIC score than $\mathcal{G}_t$, the algorithm terminates.
- $k = p$, $k = 2p$, $k = 3p$, $k = 5p$ and $k = 300$.

* Neighbour: if they can be transformed into each other by one edge addition, removal or reversal.

4. Greedy search algorithm



5. Experiments

5.1. Existing methods

- Compare three methods:
  PC-algorithm, GES, GDS.

- Evaluation:
  The structural Hamming distance between the true and estimated DAG.

* Hamming distance: this assigns a distance of 2 for each pair of reversed edges. all other edge mistakes count as 1.

5.2. Random graphs

- p=5, 20, 40.
- n=100, 500, 1000.
- β^0_{jk}: uniformly $[-1, \ -0.1] \cup [0.1, \ 1]$.
- Sparse setting: $p_{edge} = 3/(2p-2)$.
- Dense setting: $p_{edge} = 0.3$.

5.2. Random graphs

- Sparse setting

p		$n = 100$			$n = 500$			$n = 1000$		
		GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES
5	DAG	1·5	3·9	3·6	0·5	2·9	2·8	0·4	3·0	2·5
	CPDAG	1·5	2·9	2·3	0·5	1·4	1·2	0·3	1·0	0·7
20	DAG	12·2	14·1	18·0	4·5	11·1	10·3	2·7	10·1	8·7
	CPDAG	13·9	10·9	17·0	5·2	7·7	7·6	3·0	6·9	5·6
40	DAG	44·7	29·6	53·0	15·7	22·6	26·1	10·7	20·1	21·9
	CPDAG	50·0	24·4	53·1	18·9	15·9	23·4	13·4	13·3	17·5

Average structural Hamming distance

- Except for $p = 40$ and $n = 100$,
  GDS method are closer to the true DAG.

5.2. Random graphs

- Dense setting

p		$n = 100$			$n = 500$			$n = 1000$		
		GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES
5	DAG	1·2	2·9	3·0	0·6	2·4	2·2	0·3	2·1	2·1
	CPDAG	1·3	2·1	1·9	0·5	1·2	0·7	0·2	0·8	0·5
20	DAG	30·0	56·6	63·9	12·5	55·7	66·3	8·2	57·6	69·1
	CPDAG	31·0	56·1	63·2	13·1	55·5	66·2	8·8	57·5	68·5
40	DAG	216·1	242·8	323·1	185·2	247·2	430·4	172·0	248·9	470·6
	CPDAG	217·1	242·4	323·0	185·7	247·0	430·1	172·2	248·5	470·4

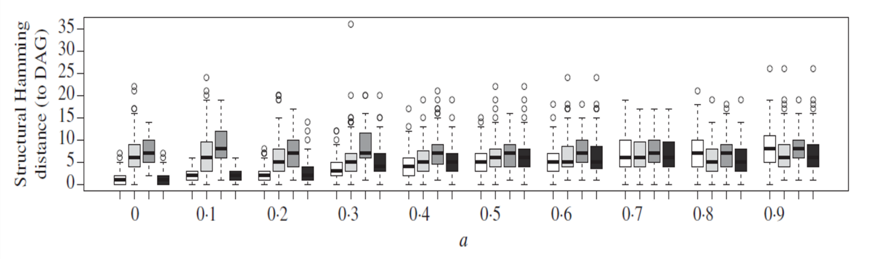
Average structural Hamming distance

- GDS method are closer to the true DAG.

5.3. Deviation from equal error variances

- $p = 10,\ n = 500$.
- σ_j^2: uniformly $[1-a,\ 1+a],\ a = (0.1, 0.2, \cdots, 0.9)$.
- p_{edge}: $2/(p-1)$ on average resulting in p edges.
- β_{jk}^0: uniformly $[-1,\ -0.1] \cup [0.1,\ 1]$.

5.3. Deviation from equal error variances



- For large values of a,

  the method does not perform worse than the

  PC-algorithm.
- GDS is relatively robust as the parameter a changes.

* Bestscore method: the result of GDS or GES depending on which

5.4. Real data

Table 3. BIC *scores of greedy equivalence search and greedy directed acyclic graph search with equal error variances obtained for different types of microarray data; smaller is better*

	Prostate	Lymphoma	Riboflavin	Leukaemia	Brain	Cancer	Colon
GES	4095	4560	2711	5456	1411	5891	3224
GDS_{EEV}	6057	5404	3236	5481	1343	6288	3201

GES, greedy equivalence search; GDS_{EEV}, greedy directed acyclic graph search with equal error variances.

- 7 datasets: Prostate, Lymphoma, Riboflavin, Leukaemia, Brain, Cancer, Colon.
- Brain, Colon datasets: GDS produced a better score than GES.

Summary

Summary

- Identifiability
  - Gaussian SEMs with the equal error variances.
  - Assumptions: Non-zero coefficient, Causal sufficiency.
- Learning
  - Penalized MLE.
  - Greedy search algorithm.

Reference

Reference

- J.Peters and P.Buhlmann, Identifiability of Gaussian Structural Equation Models with Equal Error Variances 2014.

- Park, Gunwoong, High Dimensional Gaussian DAG Model Learning via ℓ_1 - regularized Regression.


THANK YOU