Causal structure learning III

Causality

Christina Heinze-Deml

Spring 2019

Announcements

- Course evaluation

- Next week:
  - Normal lecture from 10-11
  - In-class exercise from 11-12
    - Jupyter notebook and R
    - Can also ask questions about the course material and the series

- No class on May 29th

Last week

- PC algorithm
- GES algorithm

Today

- Restricted SEMs

Score-based causal structure learning

- Constraint-based methods exploit independence statements to infer the graph
- Now: test different graph structures in their ability to fit the data
  - Address structure learning as a model selection problem

- Idea:
  - Given n i.i.d. observations $\mathcal{D} = (x_{\boldsymbol{V}}^1, \dots, x_{\boldsymbol{V}}^n)$, assign a score $\mathcal{S}(\mathcal{D}, G)$ to each graph G
  - $\mathcal{S}(\mathcal{D}, G)$ measures how well G fits the data
  - Search over the space of DAGs to find the graph with the highest score:

$$\hat{G} = \operatorname{argmax}_G \mathcal{S}(\mathcal{D}, G)$$

Score-based causal structure learning

- How to search?
  - Number of DAGs with p nodes grows super-exponentially
  - Hence, exhaustive search is often infeasible
  - Use greedy search techniques instead
    - At each step, candidate graph and set of “neighboring graphs”
    - For each neighbor, compute score
      - Take the best-scoring graph as the new candidate graph
      - If no neighbor obtains a better score, search procedure terminates
    - Result may be a local optimum only

Greedy Equivalence Search

- Assumes Markov + faithfulness
- Optimizes the BIC
- Searches over Markov equivalence classes instead of DAGs

- Greedy Equivalence Search (GES):
  - Start with empty graph
  - Two phases:
    1. Add edges until a local maximum is reached
    2. Remove edges until a local maximum is reached
  - Output is the CPDAG attaining the local maximum in phase 2

Greedy Equivalence Search

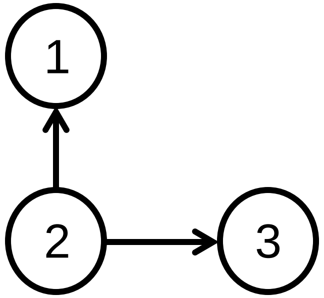
- $\mathcal{E}$: an equivalence class of DAG models

- Neighborhood relation in phase 1 of GES:
  - $\boldsymbol{\mathcal{E}}^+(\mathcal{E})$: neighbors of state $\mathcal{E}$ in phase 1
  - An equivalence class $\mathcal{E}'$ is in $\boldsymbol{\mathcal{E}}^+(\mathcal{E})$ if and only if there is some DAG $G \in \mathcal{E}$ to which we can add a single edge that results in a DAG $\mathrm{G}' \in \mathcal{E}'$

- Neighborhood relation in phase 2 of GES:
  - $\boldsymbol{\mathcal{E}}^-(\mathcal{E})$: neighbors of state $\mathcal{E}$ in phase 2
  - An equivalence class $\mathcal{E}'$ is in $\boldsymbol{\mathcal{E}}^-(\mathcal{E})$ if and only if there is some DAG $G \in \mathcal{E}$ from which we can remove a single edge that results in a DAG $\mathrm{G}' \in \mathcal{E}'$

Clicker question – Neighborhood relations in GES

- Consider the following DAG G:



- True or false?
  - The equivalence class of G, $\mathcal{E}(G)$, contains four DAGs. false
  - $\boldsymbol{\mathcal{E}}^{+}(\mathcal{E})$ contains two equivalence classes with two DAGs each. false
  - $\boldsymbol{\mathcal{E}}^{+}(\mathcal{E})$ contains three equivalence classes with two DAGs each. false
  - $\boldsymbol{\mathcal{E}}^{-}(\mathcal{E})$ contains two equivalence classes with two DAGs each. true

Structure identifiability

- Seen so far:
  - Constraint-based methods: PC, SGS
  - Score-based method: GES

- Estimand: CPDAG

Structure identifiability

- Constraint-based methods assume Markov condition + faithfulness
  - One-to-one correspondence between d-separations in G and conditional independences in P
  - Can reject all graphs outside the correct Markov equivalence class
  - Because Markov condition and faithfulness only restrict the conditional independences in P, cannot distinguish between two Markov equivalent graphs
    - I.e. under Markov + faithfulness, the Markov equivalence class (MEC) of G is identifiable from P

- Example: $X \to Y$, $X \leftarrow Y$ imply the same d-separation: $X \pm Y$

Structure identifiability

- Score-based methods assume Markov condition and a parametric model
  - E.g. linear Gaussian equations
    - Cannot distinguish between Markov equivalent graphs
    - Search is over Markov equivalence classes instead of DAGs
  - Typically faithfulness or causal minimality is assumed as well

Example: Linear models with additive Gaussian noise

- Consider the two variable case
- $X \rightarrow Y$, $X \leftarrow Y$ imply the same d-separation: $X \not\perp Y$
- Either graph can induce any distribution

Structure identifiability

- Alternative to
  - Only assuming Markov + faithfulness ← like PC
  - Or assuming Markov + faithfulness and linear Gaussian eqs. ← like GES

  use structural equation model framework and restrict function class differently

- Recall: In a SEM, each X_i is generated as a function of its graphical parents in G and noise ϵ_i:

$$X_i \leftarrow h_i\left(X_{\mathrm{pa}(i)}, \epsilon_i\right), \qquad i \in \boldsymbol{V}$$

  where $\epsilon_i, i \in \boldsymbol{V}$, are jointly independent

Structure identifiability

- Recall: In a SEM, each X_i is generated as a function of its graphical parents in G and noise ϵ_i:

$$X_i \leftarrow h_i(X_{\mathrm{pa}(i)}, \epsilon_i), \qquad i \in \boldsymbol{V}$$

  where $\epsilon_i, i \in \boldsymbol{V}$, are jointly independent

- Various options to restrict the function class
- We will refer to this class of models as restricted SEMs

Additive noise models

- We call an SEM an additive noise model (ANM) if the structural assignments are of the form

$$X_i \leftarrow h_i\left(X_{\mathrm{pa}(i)}\right) + \epsilon_i, \qquad i \in \boldsymbol{V},$$

  i.e., the noise is additive.

- Assume causal minimality; here, this means that each function $h_i(\cdot)$ is not constant in any of its arguments

- Question: Can we now obtain full structure identifiability?

Linear models with additive Gaussian noise

- Theorem: Assume that P admits the linear model

$$Y = \alpha X + N_Y, \qquad N_Y \perp\!\!\!\perp X,$$

with continuous random variables X, N_Y and Y. Then there exist $\beta \in \mathbb{R}$ and a random variable N_X such that

$$X = \beta Y + N_X, \qquad N_X \perp\!\!\!\perp Y,$$

if and only if N_Y and X are Gaussian.

- If $X = C$ and $Y = E$, it is sufficient that C or N_E are non-Gaussian to render the causal direction identifiable
  - Carries over to the multivariate case

Additive noise models

- Will look at:
  - Linear models with non-Gaussian noise
  - Nonlinear models with Gaussian noise

RESIT: Regression with subsequent independence test

- Special case of RESIT for 2 variables

1. Regress Y on X using some (possibly non-linear) regression technique
   - Denote the regression function with $\hat{f}_Y(X)$
2. Test whether $Y - \hat{f}_Y(X)$ is independent of X
3. Repeat the procedure with exchanging the roles of X and Y
4. If independence is not rejected for one direction and rejected for the other, infer the former one as the causal direction
   - In practice: Infer direction with higher *p*-value for rejecting independence as causal one

- Alternative: score-based approach
  - See Jupyter notebook next week

LiNGAM: Linear non-Gaussian acyclic models

- Linear SEM

$$X \leftarrow BX + \epsilon \quad \text{with } B \in \mathbb{R}^{p\times p}, X \in \mathbb{R}^p, \epsilon \in \mathbb{R}^p$$

- Gaussian noise
  - Cannot distinguish between Markov equivalent graphs
  - See previous example: different B matrices generate the same distribution of X

- Non-Gaussian data
  - Not fully determined by mean and covariance
  - Higher moments may contain more information

LiNGAM: Linear non-Gaussian acyclic models

- Linear SEM

$$X \leftarrow BX + \epsilon \quad \text{with } B \in \mathbb{R}^{p\times p}, X \in \mathbb{R}^p, \epsilon \in \mathbb{R}^p$$

- LiNGAM: Linear Non-Gaussian Acyclic Models
  - ϵ is mean-zero non-Gaussian with positive variance
    - Noise components are mutually independent, i.e. no hidden variables (causal sufficiency)
  - No faithfulness assumption needed
  - Estimand: DAG

LiNGAM: Linear non-Gaussian acyclic models

- Definition: Given a DAG G, we say that a bijective mapping π

$$\pi: \{1, \ldots, p\} \rightarrow \{1, \ldots, p\}$$

  is a causal ordering of the variables if it satisfies

$$\pi(i) < \pi(j) \text{ if } j \in \mathrm{desc}(i)$$

- Because of acyclicity, there is always a causal ordering
  - Not necessarily unique

$2 \rightarrow 1$

$\pi(2) = 1 \quad \pi(1) = 2$

$2 \rightarrow 1$

3

$\pi(3) = 1$
$\pi(2) = 2$
$\pi(1) = 3$

OR

$\pi(2) = 1$
$\pi(1) = 2$
$\pi(3) = 3$

$\ldots$

LiNGAM: Linear non-Gaussian acyclic models

- Linear SEM

$$X \leftarrow BX + \epsilon \quad \text{with } B \in \mathbb{R}^{p\times p}, X \in \mathbb{R}^p, \epsilon \in \mathbb{R}^p$$

- Due to acyclicity, the diagonal elements of B are zero
  - No self-loops, i.e. no edge from a node to itself

- Permuting the order of the variables using a causal ordering makes B strictly lower triangular
  - I.e. due to acyclicity, always possible to perform simultaneous, equal row and column permutations on B to make it strictly lower triangular

Example

LiNGAM: Linear non-Gaussian acyclic models

- Goal: based on n i.i.d. observations of $X_{\mathbf{V}}$ estimate B
  - Since non-zeros in B are edges in the DAG, we are also learning the DAG

- Estimation methods:
  - ICA-LiNGAM
    - Can exploit fast ICA methods but issues with local optima
  - DirectLiNGAM
    - Guaranteed convergence as $n \to \infty$

Independent component analysis

- Independent component analysis (ICA)
  - “non-Gaussian variant of factor analysis”
  - “cocktail party problem”

  - ICA model

$$X = AS$$

    - $X \in \mathbb{R}^p$: observed variables
    - $S \in \mathbb{R}^p$: mutually independent, continuous latent non-Gaussian variables – “sources”
    - $A \in \mathbb{R}^{p \times p}$: unobserved full-rank mixing matrix

  - If S is non-Gaussian, then A is identifiable up to permutation, scaling and sign of the columns

LiNGAM: Linear non-Gaussian acyclic models

- Can write:

$$X = BX + \epsilon$$

$$(I - B)X = \epsilon$$

$$X = (I - B)^{-1}\epsilon$$

- LiNGAM is an instance of the ICA model $X = AS$ with $A = (I - B)^{-1}$ and $S = \epsilon$

- Recall: A is identifiable up to permutation, scaling and sign of the columns
  - Can exploit further properties of B: “zeros on the diagonal” and “strictly lower triangular”

Example

LiNGAM: Linear non-Gaussian acyclic models

- ICA-LiNGAM algorithm:
  1. Given n i.i.d. observations of $X_{\mathbf{V}}$, use ICA to estimate $\mathrm{W} = A^{-1} = (I - B)$ up to permutation, scaling and sign of the columns
  2. Find unique permutation of the rows of W that yields $\widetilde{\mathrm{W}}$ without any zeros on the diagonal
     - Permutation is found by minimizing $\sum_i 1/|\widetilde{W}_{ii}|$ (classical linear assignment problem)
  3. Divide each row of $\widetilde{\mathrm{W}}$ by its diagonal element to yield $\widetilde{\mathrm{W}}'$ with only ones on the diagonal
  4. Compute $\hat{B} = I - \widetilde{\mathrm{W}}'$
  5. Find causal order by making $\tilde{B} = \tilde{P}\hat{B}\tilde{P}^T$ as close as possible to strictly lower triangular
     - Prune edge weights, e.g. using sparse regression

Recap

- Concepts to know:
  - Structure identifiability
    - Linear models with Gaussian noise
  - Restricted SEMs
    - Additive noise models
    - RESIT
    - LiNGAM

References and acknowledgments

- Restricted SEMs
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapters 4.1.1 - 4.1.4, 4.2.1
  - Shimizu (2014). LINGAM: Non-Gaussian Methods for estimating causal structures. Behaviormetrika.