ETH zürich

Causal structure learning II

Causality

Christina Heinze-Deml

Spring 2019

Last time

- Faithfulness
- Markov equivalence
- CPDAGs
- SGS algorithm
- PC algorithm

Today

- PC algorithm
- GES algorithm
- Restricted SEMs

PC algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the PC-algorithm of Peter Spirtes and Clark Glymour:
  - Determine the skeleton
    - No edge between i and j
      $\Leftrightarrow$
      i and j are d-separated by $\mathrm{pa}(i, G)$ or $\mathrm{pa}(j, G)$
      $\Leftrightarrow$
      i and j are d-separated by a subset $\boldsymbol{S}'$ of $\mathrm{adj}(i, G)$ or of $\mathrm{adj}(j, G)$
      $\Leftrightarrow$
      $X_i \perp\!\!\!\perp X_j | X_{\boldsymbol{S}'}$ for a subset $\boldsymbol{S}'$ of $\mathrm{adj}(i, G)$ or of $\mathrm{adj}(j, G)$
    - Start with the complete graph
    - For $k = 0, 1, \ldots, p-2$
      - Consider all pairs of adjacent vertices (i, j), and remove edge if X_i and X_j are conditionally independent given some subset of size k of $\mathrm{adj}(i, G)$ or of $\mathrm{adj}(j, G)$

PC algorithm

- Assuming Markov and faithfulness, a CPDAG can be estimated by the PC-algorithm of Peter Spirtes and Clark Glymour:
  - Determine the skeleton
  - Determine the v-structures
    - By checking for conditional dependence
  - Direct as many of the remaining edges as possible
    - By consistency with already directed edges

PC algorithm

- Assume that P has a perfect map G. Verify the following statements about the oracle version of the PC algorithm:
  - If an edge $i - j$ is removed at some point in the PC algorithm, then i and j are not adjacent in G.
  - At any point in the algorithm, the current skeleton is a supergraph of the skeleton of G.
  - If an edge $i - j$ is not removed in the PC algorithm, then i and j are adjacent in G.
  - The output of the skeleton phase of the PC algorithm is the skeleton of G.
  - [See Series 5.]

PC algorithm – sample version

- Instead of a conditional independence "oracle", we perform conditional independence tests
- In the multivariate Gaussian setting, this is equivalent to testing for zero partial correlation: $H_0 : \rho_{ij|S} = 0$ versus $H_A : \rho_{ij|S} \neq 0$

- The significance level α serves as a tuning parameter for the PC algorithm
  - Do not necessarily want to treat type I error as in traditional testing

Examples

Partial correlation

- We call X and Y uncorrelated if

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = 0$$

- We say that X and Y are partially uncorrelated given Z if

$$\rho_{X,Y|Z} = \frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Z,Y}}{\sqrt{(1-\rho_{X,Z}^2)(1-\rho_{Z,Y}^2)}} = 0$$

- $\rho_{X,Y|Z}$ equals correlation between residuals after linearly regressing X on Z and Y on Z

Partial correlation

- In general,

$$\rho_{X,Y|Z} = 0 \;\not\Rightarrow\; X \perp\!\!\!\perp Y|Z \qquad \text{and} \qquad \rho_{X,Y|Z} = 0 \;\not\Leftarrow\; X \perp\!\!\!\perp Y|Z$$

  - See Elements of Causal Inference, Example 7.9

Constraint-based causal structure learning

- Summary
  - Constraint-based methods assume Markov + 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 (MEC)
  - Estimand: CPDAG encoding the MEC

- Challenges
  - In finite samples, testing results may contradict each other
  - Choice of tuning parameter α
  - Nonparametric conditional independence tests are difficult to perform with finite samples

Constraint-based causal structure learning

- Partial identification of causal effects
  - IDA algorithm (Maathuis et al., 2009)
    - Estimate causal effects for each DAG in the Markov equivalence class
    - Infer bounds on the causal effects

- Hidden variables
  - FCI (Fast Causal Inference) algorithm and extensions (RFCI, FCI+)
    - Estimand: Partial ancestral graph (PAG)

Score-based causal structure learning

- So far: directly exploited 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 define $\mathcal{S}(\mathcal{D}, G)$?
  - Assume parametric model
    - E.g. linear Gaussian equations
    - Introduces a set of parameters θ
  - Consider the maximum likelihood estimator $\hat{\theta}$ for θ

  - Idea 1: use log-likelihood $\log p\left(\mathcal{D}\middle|\hat{\theta}, G\right)$ for the score
    - Issue: will result in the full DAG because more parameters always improve the model fit
    - Solution: use penalized log-likelihood

Score-based causal structure learning

- How to define $\mathcal{S}(\mathcal{D}, G)$?
  - Assume parametric model
    - E.g. linear Gaussian equations
    - Introduces a set of parameters θ
  - Consider the maximum likelihood estimator $\hat{\theta}$ for θ

  - Idea 2: use the Bayesian Information Criterion (BIC) for the score

$$\mathcal{S}(\mathcal{D}, G) = 2 \log p\left(\mathcal{D} \middle| \hat{\theta}, G\right) - \log n \cdot (\#\text{parameters})$$

    - $\log p\left(\mathcal{D} \middle| \hat{\theta}, G\right)$ is the log-likelihood
    - Second term: complexity penalty for number of parameters
  - Other possibilities exist

Score-based causal structure learning

- Properties of the BIC
  - Score-equivalence: Score is the same for all DAGs within the same MEC
  - Decomposable: Score can be computed as a sum of terms each of which is a function only of one node and its parents

$$\mathcal{S}(\mathcal{D}, G) = \sum_j s(X_j, X_{\mathrm{pa}(j)})$$

    - Computational savings
  - Local consistency: Score of a DAG model G …
    - … increases if adding an edge that eliminates an independence constraint that does not hold in P
    - … decreases if adding an edge that does not eliminate such a constraint

Example

Score-based causal structure learning

- How to search?
  - Recall from week 2 – Number of DAGs with p nodes:

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	9942119532215951589522891459235452451655502687858830501478 3
19	332771901227107591736177573311261125883583076258421902583546773505
20	2344880451051088988152559855229099188899081192234291298795803236068491263

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

Score-based causal structure learning

- 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

- Phase 1: Hope is to identify a model that is as simple as possible
  - Ignoring computational cost, one could also start with a full DAG and only do phase 2

Score-based causal structure learning

- 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

- Chickering (2002): GES identifies the true CPDAG as $n \to \infty$
- GES can be improved by including a "turning phase" (Hauser and Bühlmann (2012))

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: $X \perp\!\!\!\perp Y$

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
- $X \to Y$, $X \leftarrow Y$ imply the same d-separation: $X \perp\!\!\!\perp Y$
- Either graph can induce any distribution

Structure identifiability

- Alternative to
  - Only assuming Markov + faithfulness
  - Or assuming Markov + faithfulness and linear Gaussian eqs.

  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:

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

  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

- Two variable case

- 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.

Recap

- Concepts to know:
  - Partial correlation and conditional independence
  - Score-based causal structure learning
    - GES
  - Structure identifiability
    - Linear models with Gaussian noise
  - Restricted SEMs
    - Additive noise models

References and acknowledgments

- Score-based structure learning
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapter 7.2.2
  - (Optional) Chickering (2002). Optimal Structure Identification with Greedy Search. JMLR.
  - (Optional) Hauser and Bühlmann (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. JMLR.

- Restricted SEMs
  - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapters 4.1.1 - 4.1.4, 4.2.1