The (Casual) Causality Course
Introduction, some notations and fundamentals

Gherardo Varando

IPL

02 May 2023

The course

- **When:** Tuesday and Thursday from 10:30 to 13:00 for the next 4 weeks
- **Where:** In person IPL hall and (limited) virtual at zoom link `https://uv-es.zoom.us/j/94222318081?pwd=NkNGOForRDYrNEtuUHR2cGFmQWN6Zz09`
- **Who:** Gherardo Varando (gherardo.varando@uv.es) and Emiliano Díaz (emiliano.diaz@uv.es)
- **Material** available on github `https://github.com/IPL-UV/casual_causality_course` if you are in the IPL you should have access to the private repo, otherwise write me an email with your github email and username and I will grant you access.

Content

Week 1 **Introduction and notations**

Tue 02 Course introduction and first concepts

Thur 04 Causal frameworks and framing causal problems

Week 2 **Causal Discovery**

Tue 09 Classical approaches

Thur 11 Continuous optimization methods and NN parametrizations

Week 3 **Causal Inference**

Tue 16 Causal effect estimation and possible biases

Thur 18 Machine learning methods for causal effect estimation

Week 4 **Applications**

Tue 23

Thur 25

Content

Week 1 **Introduction and notations**

- Tue 02 Course introduction and first concepts
- Thur 04 Causal frameworks and framing causal problems

Week 2 **Causal Discovery**

- Tue 09 Classical approaches
- Wed 10? Practical session on causal discovery?
- Thur 11 Continuous optimization methods and NN parametrizations

Week 3 **Causal Inference**

- Tue 16 Causal effect estimation and possible biases
- Thur 18 Machine learning methods for causal effect estimation

Week 4 **Applications**

- Tue 23
- Thur 25

Statistical Models: a very short story

- A mathematical model of the data generating process
- A description of the probability of data
- A collection of statistical assumptions
- Allows us to compute probabilities of events

Definition

Formally we can define a statistical model as a pair $(\mathcal{S}, \mathcal{P})$ where

- $\mathcal{S}$ is a sample space, formally $\mathcal{S} = (\Omega, \mathcal{F})$ with $\mathcal{F}$ a σ-algebra on Ω a set
- $\mathcal{P}$ is a collection of probability distributions on $\mathcal{S}$

Usually $\mathcal{P}$ is indexed by some finite-dimensional parameter θ, in that case the model is said to be **parametric**, if instead the parameter is infinite dimensional (or there is no parameter) the model is said to be **non-parametric**.

[Wasserman, 2004]

Inference and statistical problems

- modelling the data

Inference and statistical problems

- modelling the data
- prediction or forecasting

Inference and statistical problems

- modelling the data
- prediction or forecasting
- queries on parameter(s)

Inference and statistical problems

- modelling the data
- prediction or forecasting
- queries on parameter(s)
- decision making

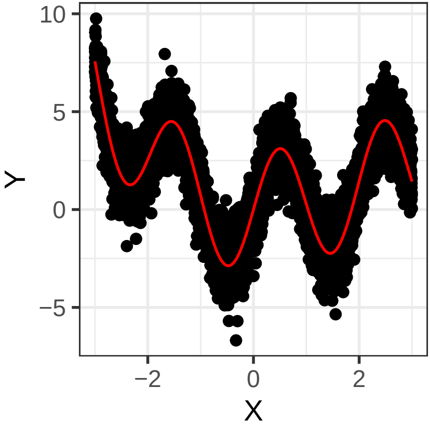
Inference and statistical problems

- modelling the data
- prediction or forecasting
- queries on parameter(s)
- decision making
- model selection

Regression models

Given data (X_i, Y_i)

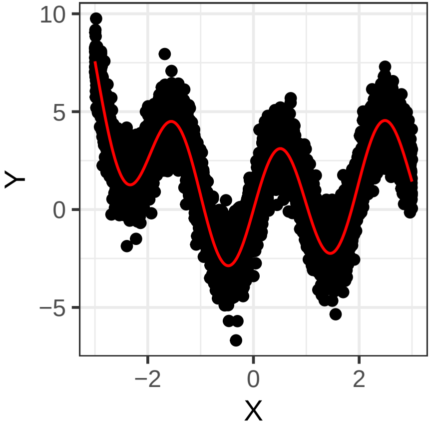
- find best function that approximate $Y = f(X)$



Regression models

Given data (X_i, Y_i)

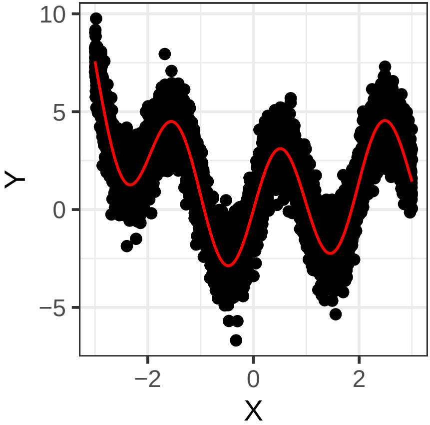
- find best function that approximate $Y = f(X)$
- define a statistical model for $P(Y|X)$, e.g. if consider an additive noise model $Y = f(X) + \varepsilon$



Regression models

Given data (X_i, Y_i)

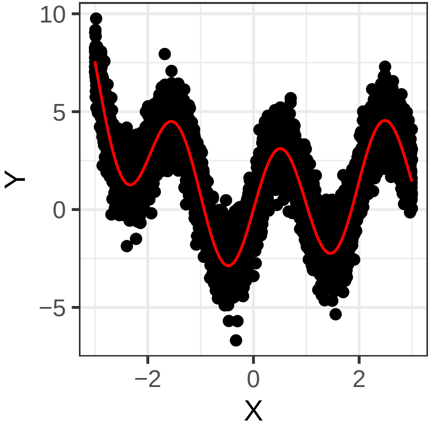
- find best function that approximate $Y = f(X)$
- define a statistical model for $P(Y|X)$, e.g. if consider an additive noise model $Y = f(X) + \varepsilon$
- linear case $Y = \sum \beta_i X_i + \beta_0 + \varepsilon$ we can do inference on the parameters β_i: confidence intervals, hypothesis testing



Regression models

Given data (X_i, Y_i)

- find best function that approximate $Y = f(X)$
- define a statistical model for $P(Y|X)$, e.g. if consider an additive noise model $Y = f(X) + \varepsilon$
- linear case $Y = \sum \beta_i X_i + \beta_0 + \varepsilon$ we can do inference on the parameters β_i: confidence intervals, hypothesis testing
- non-parametric approaches, for example kernel methods
- non-additive noise models, $Y = f(X, \varepsilon)$

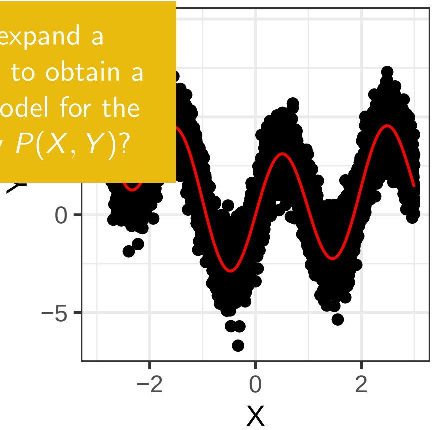


Regression models

Given data (X_i, Y_i)

- find best function that approximate Y ... $f(X)$
- define a statist... $P(Y|X)$, e.g. i... additive noise ... $Y = f(X) + \varepsilon$
- linear case $Y = \sum \beta_i X_i + \beta_0 + \varepsilon$ we can do inference on the parameters β_i: confidence intervals, hypothesis testing
- non-parametric approaches, for example kernel methods
- non-additive noise models, $Y = f(X, \varepsilon)$

How we can expand a regression model to obtain a full statistical model for the joint probability $P(X, Y)$?



Bayesian Networks

Definition

A Bayesian Network (BN) over random variables $X_1, X_2, \ldots, X_p$ is a pair (G, P) where

- G is a DAG over p nodes (indexed as the r.vs)
- P is a joint probability over $X_1, \ldots, X_p$ such that $P = \prod_{i=1}^{p} P(X_i | X_{pa(i)})$

- BNs are an example of probabilistic graphical models
- define statistical models

[Lauritzen, 1996, Koller and Friedman, 2009, Maathuis et al., 2018]

Bayesian Networks

Definition

A Bayesian Network (BN) over random variables $X_1, X_2, \ldots, X_p$ is a pair (G, P) where

- G is a DAG over p nodes (indexed as the r.vs)
- P is a joint probability over $X_1, \ldots, X_p$ such that $P = \prod_{i=1}^{p} P(X_i | X_{pa(i)})$

- BNs are an example of probabilistic graphical models
- define statistical models
- for categorical r.v.s $P(X_i | X_{pa(i)})$ can be represented with conditional probability tables (CPTs)

[Lauritzen, 1996, Koller and Friedman, 2009, Maathuis et al., 2018]

Bayesian Networks

Definition

A Bayesian Network (BN) over random variables $X_1, X_2, \ldots, X_p$ is a pair (G, P) where

- G is a DAG over p nodes (indexed as the r.vs)
- P is a joint probability over $X_1, \ldots, X_p$ such that $P = \prod_{i=1}^{p} P(X_i | X_{pa(i)})$

- BNs are an example of probabilistic graphical models
- define statistical models
- for categorical r.v.s $P(X_i | X_{pa(i)})$ can be represented with conditional probability tables (CPTs)
- for Gaussian r.v.s (and linear relationships), it is sufficient to specify a set of (linear) regressions for each node over its parents and the variance of the residuals

---

[Lauritzen, 1996, Koller and Friedman, 2009, Maathuis et al., 2018]

Bayesian Networks

Definition

A Bayesian Network (BN) over random variables $X_1, X_2, \ldots, X_p$ is a pair (G, P) whe

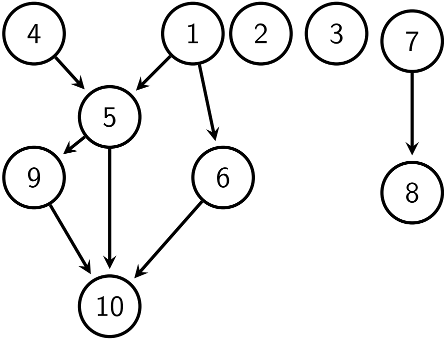
- G is a DAG)
- P is a joint hat
  $P = \prod_{i=1}^{p} P$

Some people [Wasserman, 2004] think that the term Bayesian Network is misleading and poor terminology since BN do not have anything to do with Bayesian methods

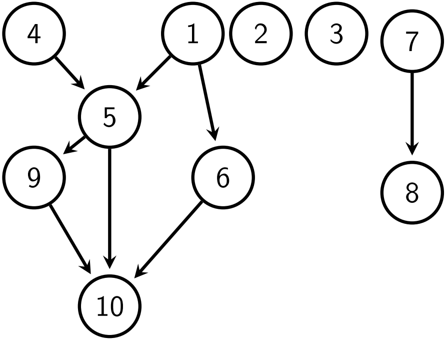
- BNs are an e models
- define statist
- for categorical r.v.s $P(X_i | X_{pa(i)})$ can be represented with conditional probability tables (CPTs)
- for Gaussian r.v.s (and linear relationships), it is sufficient to specify a set of (linear) regressions for each node over its parents and the variance of the residuals

---

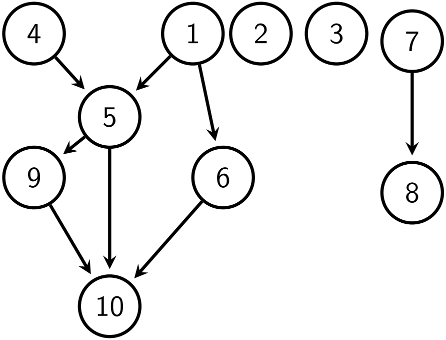
[Lauritzen, 1996, Koller and Friedman, 2009, Maathuis et al., 2018]


4
1
2
3
7
5
9
6
8
10

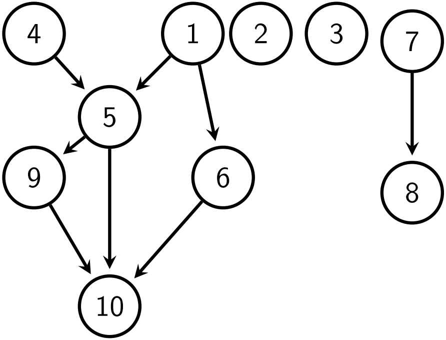
- 1. parents $pa(i)$
  2. children $ch(i)$
  3. descendants $de(i)$
  4. non-descendants $nde(i)$
  5. ancestrors $an(i)$



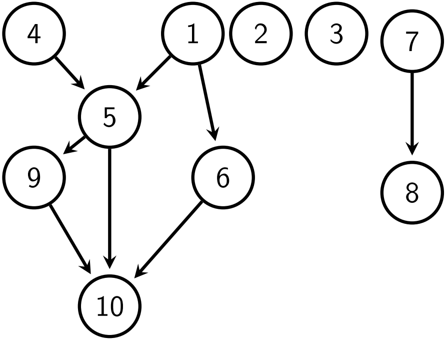
- 1. parents $pa(i)$
  2. children $ch(i)$
  3. descendants $de(i)$
  4. non-descendants $nde(i)$
  5. ancestrors $an(i)$
- v-structures, immoralities



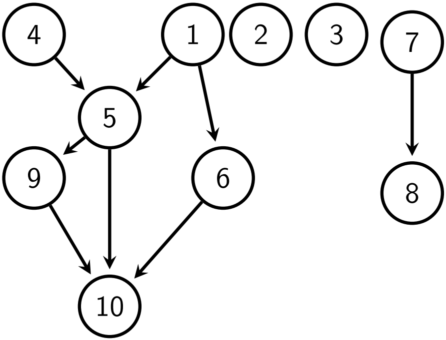
- 1. parents $pa(i)$
  2. children $ch(i)$
  3. descendants $de(i)$
  4. non-descendants $nde(i)$
  5. ancestrors $an(i)$
- v-structures, immoralities
- moral graph G^m



- 1. parents $pa(i)$
  2. children $ch(i)$
  3. descendants $de(i)$
  4. non-descendants $nde(i)$
  5. ancestrors $an(i)$
- v-structures, immoralities
- moral graph G^m
- topological order

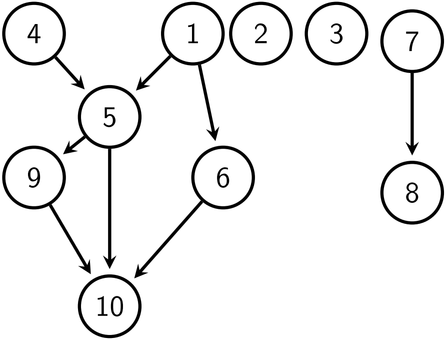


- 1. parents $pa(i)$
  2. children $ch(i)$
  3. descendants $de(i)$
  4. non-descendants $nde(i)$
  5. ancestrors $an(i)$
- v-structures, immoralities
- moral graph G^m
- topological order



- 1
  2
  3 $i)$
  4 $nde(i)$
  5
- v-s alities
- mo
- top

can you write the factorization of the joint probability associated with this graph?



- BNs are statistical models

BN/DAG and conditional independences

The following statements are equivalent:

- (G, P) is a BN, that is P factorize recursively wrt the DAG G
- P satisfies the *local Markov property* wrt G, that is $X_i \perp\!\!\!\perp X_{nd(i)} | X_{pa(i)}$
- P satisfies the *global Markov property* wrt G, that is $X_A \perp\!\!\!\perp X_B | X_D$ whenever A and B are d-separated by D in DAG G (A and B are separated by D in $G^m_{an(A \cup B \cup D)}$)

- BNs are statistical models
- statistical models are "collection of statistical assumptions"

BN/DAG and conditional independences

The following statements are equivalent:

- (G, P) is a BN, that is P factorize recursively wrt the DAG G
- P satisfies the *local Markov property* wrt G, that is $X_i \perp\!\!\!\perp X_{nd(i)} | X_{pa(i)}$
- P satisfies the *global Markov property* wrt G, that is $X_A \perp\!\!\!\perp X_B | X_D$ whenever A and B are d-separated by D in DAG G (A and B are separated by D in $G^m_{an(A \cup B \cup D)}$)

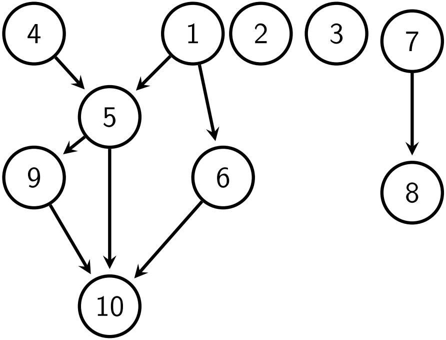
- BNs are statistical models
- statistical models are "collection of statistical assumptions"
- which are the assumptions associated with a given BN (G, P)?

BN/DAG and conditional independences

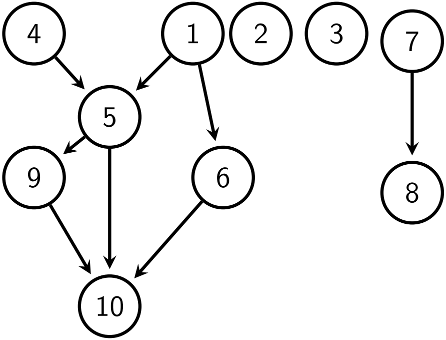
The following statements are equivalent:

- (G, P) is a BN, that is P factorize recursively wrt the DAG G
- P satisfies the *local Markov property* wrt G, that is $X_i \perp\!\!\!\perp X_{nd(i)} | X_{pa(i)}$
- P satisfies the *global Markov property* wrt G, that is $X_A \perp\!\!\!\perp X_B | X_D$ whenever A and B are d-separated by D in DAG G (A and B are separated by D in $G^m_{an(A \cup B \cup D)}$)

- d-separation, equivalent to separation in the moral graph of the ancestors of involved vertices



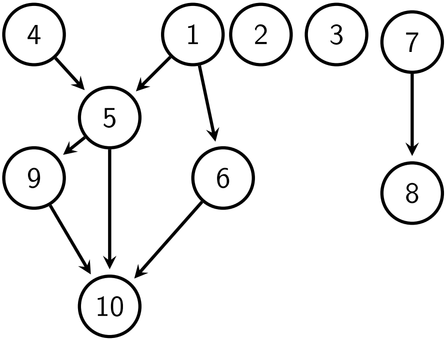
- d-separation, equivalent to separation in the moral graph of the ancestors of involved vertices
- equivalently i and j are d-separated by D if there exists no undirected path u between i and j such that
    1. every collider in u has a descendants in D
    2. no other vertex on u is in D



- d-separation, equivalent to sep… …oral gra… …ors of inv…

Can you list some d-separations in the graph? some are easy ...

- equ… …are d-s… …there exi… …path u bet… …h that
  1. every collider in u has a descendants in D
  2. no other vertex on u is in D



Sampling from a BN

How can we obtain samples from a probability distribution associated with a BN?

Sampling from a BN

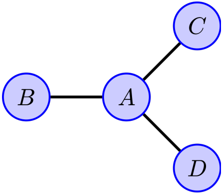
How can we obtain samples from a probability distribution associated with a BN?

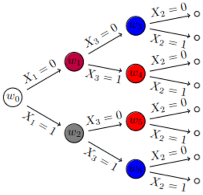
we can use the topological order to sample efficiently

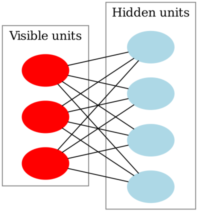
- pick a topological order of the nodes in G
- to generate each sample:
  1. start by sampling x_i from $P(X_i)$ for each nodes i without parents (there must be at least one)
  2. follow the topological order and sample from $P(X_i|X_{pa(i)} = x_{pa(i)})$ (since we follow the topological order $x_{pa(i)}$ is already sampled)

Other graphical models

- Markov networks, Markov random fields or undirected graphical models (e.g. Ising models in statistical physics) [Koller and Friedman, 2009, Lauritzen, 1996]
- Model based on event trees such as staged event trees or chain event graphs [Leonelli and Varando, 2023]
- Chain graphs
- restricted Boltzman machines







Causality?

- statistical models (and ML models) represents association in the data

Causality?

- statistical models (and ML models) represents association in the data
- stat/ML models are not mechanistic/physical/causal models of the data

Causality?

- statistical models (and ML models) represents association in the data
- stat/ML models are not mechanistic/physical/causal models of the data
- most of the time (in sciences) we are actually interested in causal questions

Causality?

- statistical models (and ML models) represents association in the data
- stat/ML models are not mechanistic/physical/causal models of the data
- most of the time (in sciences) we are actually interested in causal questions
- but what is *causality* ?

(Probabilistic) Causal Models

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Manuele Leonelli and Gherardo Varando. Context-specific causal discovery for categorical data using staged trees. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8871–8888. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/leonelli23a.html.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. CRC Press, 2018.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.