Interpretable and Robust Statistical Machine Learning

Fall 2024

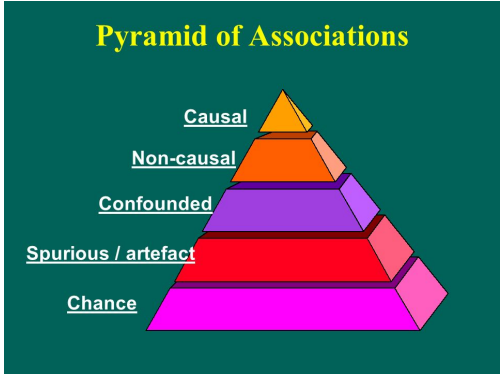
Peter Bühlmann

Lecture 2: Graphical and Causal Models

Causality

“Felix, qui potuit rerum cognoscere causas”
Fortunate who was able to know the causes of things
(Georgics, Virgil, 29 BC)

already people in ancient times (Egyptians, Greeks, Romans, Chinese) have debated on causality

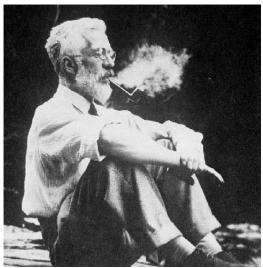


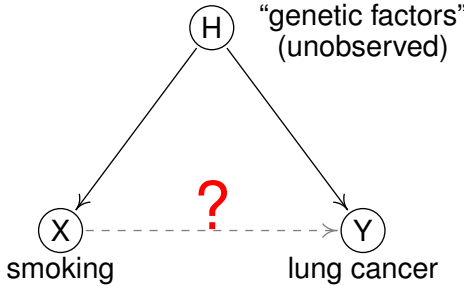
the word “causal” is very ambitious...

perhaps too ambitious...
but we aim at least at doing something “more suitable” than standard regression or classification

Recap last week: confounding is also (mostly) a causal concept

Does smoking cause lung cancer?





as a warm-up exercise...

$$\text{correlation} \neq \text{causation}$$

number of Nobel prizes vs. chocolate consumption

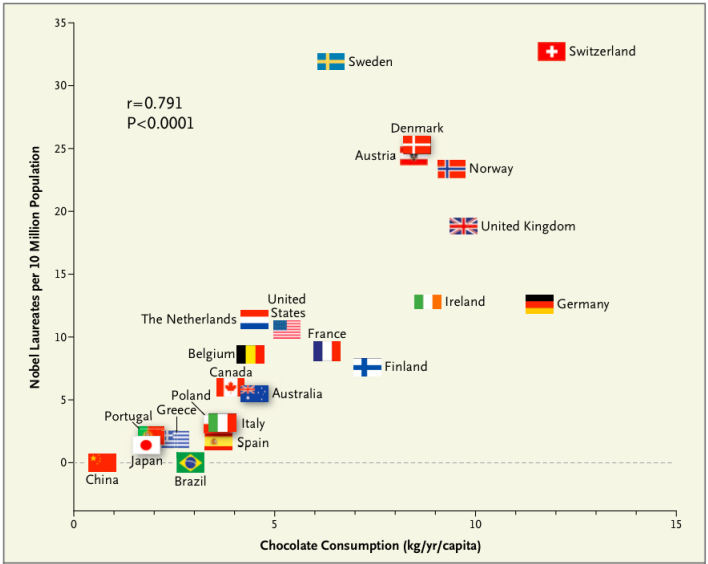


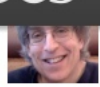
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

B

Eating chocolate produces Nobel prize winners, says study

By Oliver Nieburg, 11-Oct-2012

Related tags: noble prize, nobel laureate, Einstein, Marie Curie, chocolate, brain, Switzerland, Sweden, candy



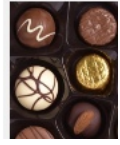
I'm a medical journalist covering cardiology news.

+ Follow (73)

PHARMA & HEALTHCARE | 10/10/2012 @ 5:02PM | 14,700 views

Chocolate And Nobel Prizes In Study

4 comments, 2 called-out + Comment Now + Follow Comments



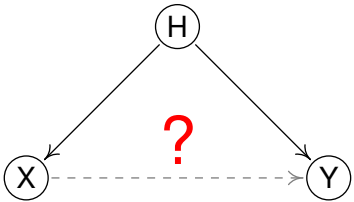
You don't have to be a genius to like chocolate, but geniuses are more likely to eat lots of chocolate, at least according to a new paper published in the August *New England Journal of Medicine*. Franz Messerli reports a highly

Possible interpretations

X: chocolate consumption; Y: obtaining Nobel prize

chocolate produces Nobel prize

geniuses eat more chocolate



hidden confounder $H =$ "wealth"

well... you might have your own theories...

well... you might have your own theories...

it would be most helpful to do:

- an experiment
- a randomized controlled trial (RCT)

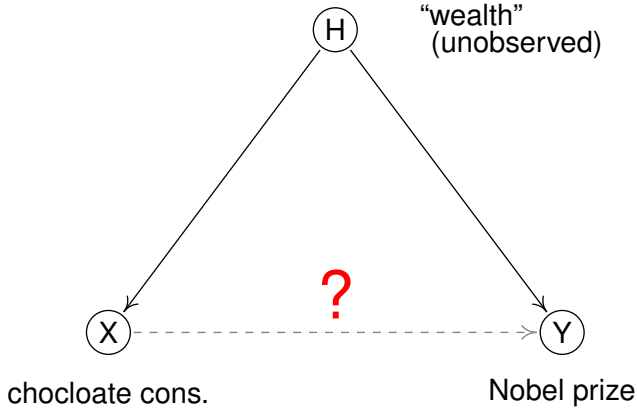
(often considered as) the gold-standard

forcing some people to eat lots and lots of chocolate!

gold-standard: a randomized controlled trial (RCT)

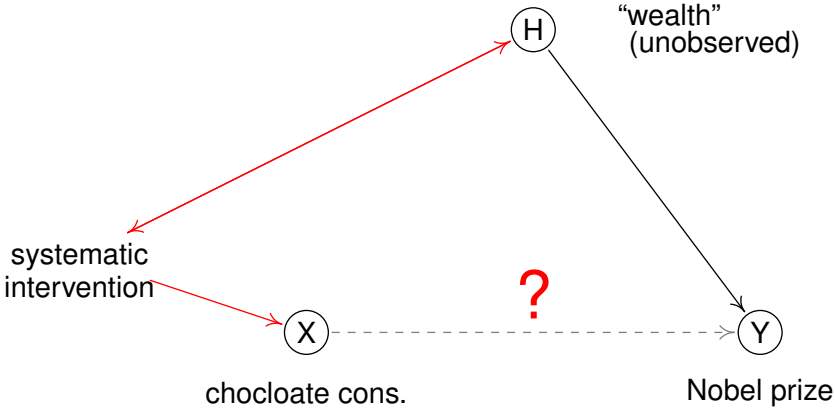
- two groups at random
  (at random: to break dependencies to hidden variables)
- force one group to eat lots of chocolate
- ban the other group from eating chocolate at all
- wait a lifetime to see what happens; and compare!

the hidden confounder is the problematic case



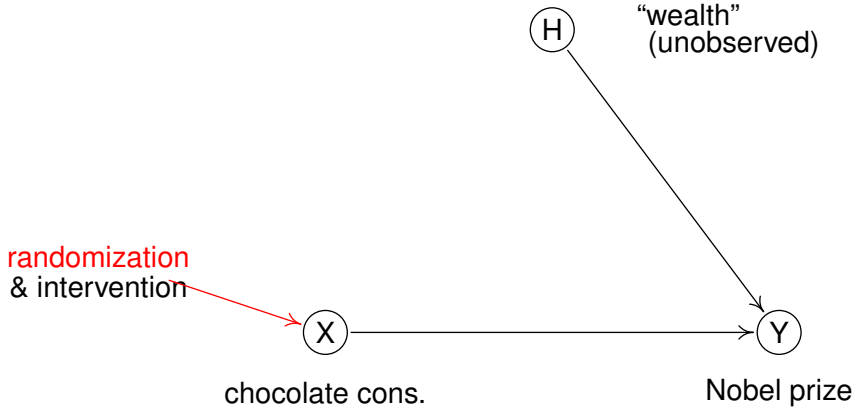
Why randomization

the hidden confounder is the problematic case



Why randomization

the hidden confounder is the problematic case



Aspects of the history

C. Peirce (1896), Fisher (1918), Neyman (1923), Fisher (1925), Holland, Rubin, Pearl, Spirtes–Glymour–Scheines, Dawid, Robins, Bollen, ...

developed in different fields including economics, psychometrics, social sciences, statistics, computer science, ...

Problems with randomized control trials (RCTs)

- randomization can be unethical
- long time horizon & reliability of participants (“non-compliance”)
- high costs
- ...

What can we say without RCTs?

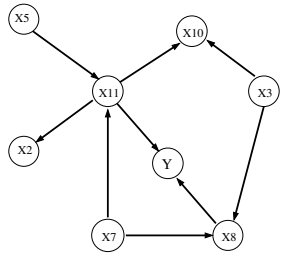
it will never be fully confirmatory
Fisher's argument on "smoking and lung cancer"

What can we say without RCTs?

in some sense, this is the main topic of the lectures!

Graphical models: a fraction of the basics

consider a directed acyclic graph (DAG) D:



$Y = X_p$

- nodes or vertices $v \in \mathcal{V} = \{1, \ldots, p\}$
- edges $e \in \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$

we identify the nodes with random variables X_v, $v = 1, \ldots, p$ (often using the index "j" instead of "v")

the edges encode "some sort of conditional dependence"

Recursive factorization and Markov properties

consider a DAG D

a distribution P of $X_1, \ldots, X_p$ allows a recursive factorization w.r.t. D if:

- P has a density $p(.)$ w.r.t. μ;
- $p(x) = \prod_{j=1}^{p} p(x_j | x_{\mathrm{pa}(j)})$,
  where $\mathrm{pa}(j)$ denotes the parental nodes of j

this factorization is intrinsically related to Markov properties:
if P admits a recursive factorization according to D:
the local Markov property holds:

$$p(x_j | x_{\setminus j}) = p(x_j | \underbrace{x_{\partial j}}_{\text{the "boundary values"}})$$

and often one simplifies and says that "P is Markovian w.r.t. D"

Recursive factorization and Markov properties

consider a DAG D

a distribution P of $X_1, \ldots, X_p$ allows a recursive factorization w.r.t. D if:

- P has a density $p(.)$ w.r.t. μ;
- $p(x) = \prod_{j=1}^{p} p(x_j | x_{\mathrm{pa}(j)})$,
  where $\mathrm{pa}(j)$ denotes the parental nodes of j

this factorization is intrinsically related to Markov properties:
if P admits a recursive factorization according to D:
the local Markov property holds:

$$p(x_j | x_{\setminus j}) = p(x_j | \underbrace{x_{\partial j}}_{\text{the "boundary values"}})$$

and often one simplifies and says that "P is Markovian w.r.t. D"

if P has a positive density $p(.)$ with respect to a product measure μ on

$$\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_p, \quad X_j \in \mathcal{X}_j \ (j = 1, \ldots, p)$$

all the global, local and pairwise Markov properties (in the corresponding undirected graphs) coincide (Lauritzen, 1996)

Global Markov property:
if C separates A and B, then

d-separation for DAGs

$$X_A \text{ independent } X_B | X_C$$

d-separation:
d-SEPARATION WITHOUT TEARS
(At the request of many readers)
`http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html`

“d-separation is a criterion for deciding, from a given DAG, whether a set X of variables is independent of another set Y, given a third set Z. The idea is to associate ”dependence” with ”connectedness” (i.e., the existence of a connecting path) and ”independence” with ”unconnectedness” or ”separation”. The only twist on this simple idea is to define what we mean by ”connecting path”, given that we are dealing with a system of directed arrows...”

Consequences

Assume that P factorizes according to D and fulfills the global Markov property ("P is Markov w.r.t. D")

Then: if A and B are d-separated in the graph D by a set $C \Longrightarrow X_A \perp X_B | X_C$

we can read off some conditional dependencies from the graph D

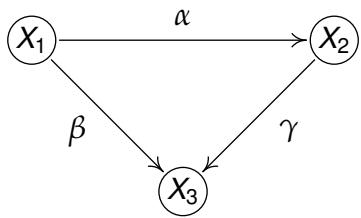
but typically not all conditional dependencies are encoded in the graph

all conditional dependencies are encoded in the graph

A distribution P is faithful w.r.t. DAG D if:

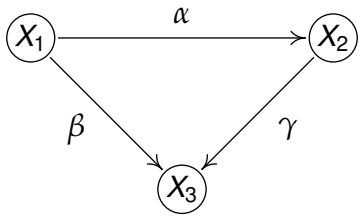
1. P is global Markov w.r.t. D
2. all conditional dependencies are encoded (by some rules which are consistent with the Markov property) from the graph D

example of a non-faithful distribution P w.r.t. a DAG D



$$
\begin{aligned}
X_1 &\leftarrow \varepsilon_1, \\
X_2 &\leftarrow \alpha X_1 + \varepsilon_2, \\
X_3 &\leftarrow \beta X_1 + \gamma X_2 + \varepsilon_3, \\
&\varepsilon_1, \varepsilon_2, \varepsilon_3 \text{ i.i.d. } \mathcal{N}(0,1)
\end{aligned}
$$

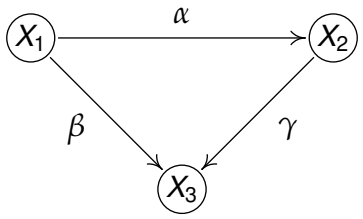
$\rightsquigarrow X_1, X_2, X_3$ jointly Gaussian



for $\beta + \alpha\gamma = 0$: $\mathrm{Corr}(X_1, X_3) = 0$; that is: $X_1 \perp X_3$

but this independence cannot be read-off from the graph by some separation rule

non-faithfulness “typically” happens by cancellation of coefficients (in linear systems)



for $\beta + \alpha\gamma = 0$: $\mathrm{Corr}(X_1, X_3) = 0$; that is: $X_1 \perp X_3$

but this independence cannot be read-off from the graph by some separation rule

non-faithfulness “typically” happens by cancellation of coefficients (in linear systems)

fact: if edge weights are sampled i.i.d. from an absolutely continuous distribution
$\rightsquigarrow$ non-faithful distributions have Lebesgue measure zero

(i.e. they are "unlikely")

but this reasoning is "statistically not valid": with finite samples, we cannot distinguish between zero correlations and correlations of order of magnitude $1/\sqrt{n}$ (and analogous for "near cancellation being of order $1/\sqrt{n}$")

$\rightsquigarrow$ the volume (the probability) of near cancellation when edge weights are sampled i.i.d. from an absolutely continuous distribution is large! Uhler, Raskutti, PB and Yu (2013)

strong faithfulness:
for $\rho(i,j|S) = \mathrm{Parcorr}(X_i, X_j|X_S)$, require:

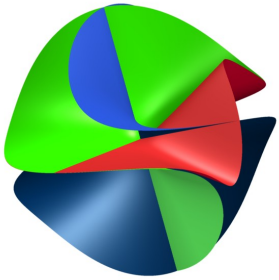
$$A(\tau, d): \qquad \min\left\{|\rho(i,j|S)|;\ \rho(i,j|S) \neq 0,\ i \neq j,\ |S| \leq d\right\} \geq \tau$$

(typically: $\tau \asymp \sqrt{\log(p)/n}$)

strong faithfulness can be rather severe

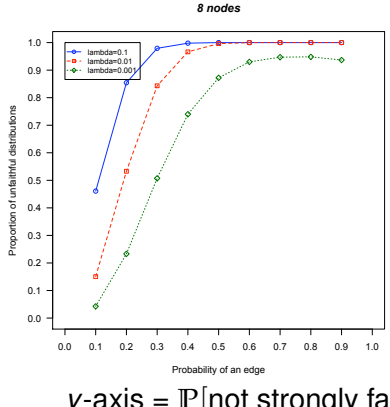
(Uhler, Raskutti, PB & Yu, 2013)

3 nodes, full graph



unfaithful distributions
due to exact cancellation

8 nodes, varying sparsity



y-axis = $\mathbb{P}[\text{not strongly faithful}]$

Consequences:
we later want to learn graphs or equivalence classes of graphs from data

when doing so via estimated conditional dependencies one needs some sort of faithfulness assumption...

Structural learning/estimation of directed graphs

motivation: directed graphs encode some "causal structure"

in a DAG:
a directed arrow $X \to Y$ says that "X is a direct cause of Y"
and we will discuss more later

goal: estimate "the true underlying DAG" from data
$\rightsquigarrow$ impossible (in general) with observational data

more precisely:

- "true" DAG D
- data-generating distribution P which allows recursive factorization w.r.t. D
- n i.i.d. data/copies of $X_1, \ldots, X_p \sim P$: $X^{(1)}, \ldots, X^{(n)}$
  the data is called "observational data": it is sampled from P and there are no interventions/perturbations involved (see later)

severe issue of identifiability: given P (or an infinite amount of data), there are several DAGs, say $D \neq D'$ such that P allows recursive factorization w.r.t. D and D'
$\rightsquigarrow$ cannot learn the true DAG D from observational data

but we can learn the "true" equivalence class of DAGs

more precisely:

- "true" DAG D
- data-generating distribution P which allows recursive factorization w.r.t. D
- n i.i.d. data/copies of $X_1, \ldots, X_p \sim P$: $X^{(1)}, \ldots, X^{(n)}$
  the data is called "observational data": it is sampled from P and there are no interventions/perturbations involved (see later)

severe issue of identifiability: given P (or an infinite amount of data), there are several DAGs, say $D \neq D'$ such that P allows recursive factorization w.r.t. D and D'
$\rightsquigarrow$ cannot learn the true DAG D from observational data

but we can learn the "true" equivalence class of DAGs

Minimal I-MAP

the statistical view:
data generating distribution P

consider the class of DAGs

$$\mathcal{D}_{\mathrm{I-MAP}}(P) = \{\text{DAG } D;\ \underbrace{P \text{ allows rec. factor. w.r.t. } D}_{P \text{ "is Markovian w.r.t. } D\text{"}}\}$$

$$\mathcal{D}_{\text{minimal I-MAP}}(P) = \{D \in \mathcal{D}_{\mathrm{I-MAP}}(P);\quad \underbrace{|D| = \min_{D' \in \mathcal{D}_{\mathrm{I-MAP}}(P)} |D'|}_{D \text{ has minimal no. of edges}}\ \}$$

in my opinion: this is the most natural definition for statistical purposes... (van de Geer & PB, 2013)
... since we start with the data generating distribution

Markov equivalence class

the much more common (and more complicated?) definition
consider

$$\mathcal{M} = \{\text{positive densities on } \underbrace{\mathcal{X}}_{\text{support of } X_1, \ldots, X_p}\}$$

for a DAG D:

$$\mathcal{M}(D) = \{p \in \mathcal{M};\ p \text{ allows rec. fact. w.r.t. } D\}$$

DAGs D and D' are Markov equivalent if $\mathcal{M}(D) = \mathcal{M}(D')$:
write $D \sim D'$

equivalence relation leads to
Markov equivalence class $\mathcal{D}_{\text{Markov}}(D)$ for a DAG D

note that Markov equivalence involves consideration of many distributions; not just the data generating distribution
("usual language in graphical modeling")
Markov equiv. "starts" from a DAG D (e.g. the "true causal DAG")

consider true underlying DAG D^0 (for causality, this will be important – see later) and data generating distribution P which is faithful w.r.t. D^0

then:

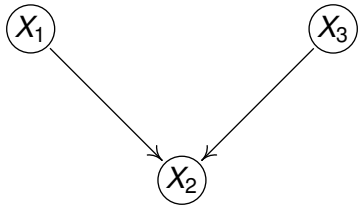
$$\mathcal{D}_{\text{minimal I-MAP}}(P) = \mathcal{D}_{\text{Markov}}(D^0)$$

Theorem (Verma & Pearl, 1990)
Two DAGs D and D' are Markov equivalent if and only if

- they have the same skeleton (undirected graph removing edge directions)
- they have the same v-structures

a graphical criterion only!

v-structure



Markov equivalence class:

An equivalence class can be uniquely represented by a completed partially directed acyclic graph (CPDAG)

Structural learning algorithms (in high dimensions)

for Markov equivalence class or class of minimal I-MAPs

most popular:

- constraint-based
  relying on inferring conditional dependencies
  $\rightsquigarrow$ requires strong faithfulness assumption

  PC-algorithm (**P**eter Spirtes & **C**lark Glymour, 1991)
- score-based methods
  in particular penalized Gaussian likelihood
  no faithfulness assumption for class of minimal I-MAPs

  GES-algorithm: Greedy Equivalence Search (Chickering, 2002)

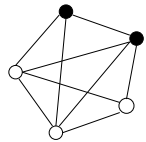
The PC-algorithm (Spirtes & Glymour, 1991)

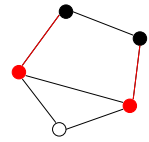
- crucial assumption:
  distribution P (strongly) faithful to the true underlying DAG
- less crucial but convenient:
  Gaussian assumption for $X_1, \ldots, X_p \rightsquigarrow$ can work with partial correlations for inferring conditional dependencies

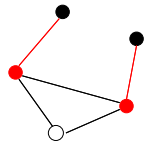
- input: $\hat{\Sigma}_{\mathrm{MLE}}$
  but we only need to consider many small sub-matrices of it (assuming sparsity of the graph)
- output: based on a clever data-dependent (random) sequence of multiple tests
  estimated CPDAG (i.e., Markov equivalence class)

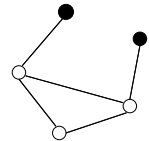
PC-algorithm: a rough outline
for estimating the skeleton of underlying DAG

1. start with full graph
2. remove edge $i - j$ if $\widehat{\text{Cor}}(X_i, X_j)$ is small
   (Fisher's Z-transform and null-distribution of zero correlation)
3. partial correlations of order 1:
   remove edge $i - j$ if $\widehat{\text{Parcor}}(X_i, X_j | X_k)$ is small for some k in the current neighborhood of i or j (thanks to faithfulness)

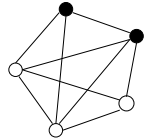

full graph

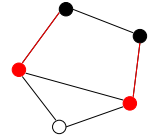

correlation screening

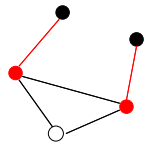

partial correlation order 1

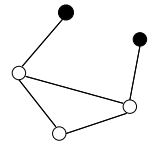

stopped

4. move-up to partial correlations of order 2:
   remove edge $i - j$ if partial correlation $\widehat{\text{Parcor}}(X_i, X_j | X_k, X_\ell)$ is small for some k, ℓ in the current neighborhood of i or j (thanks to faithfulness)
5. until removal of edges is not possible anymore,
   i.e. stop at minimal order of partial correlation where edge-removal becomes impossible


full graph


correlation screening


partial correlation order 1


stopped

additional step of the algorithm needed for estimating directions yields an estimate of the CPDAG (equivalence class of DAGs)
R-package: `pcalg` (Kalisch et al., 2012)

Statistical theory (Kalisch & PB, 2007)

n i.i.d. observational data points; p variables
high-dimensional setting where $p \gg n$

assumptions:

- $X_1, \ldots, X_p \sim \mathcal{N}_p(0, \Sigma)$ Markov and faithful to true DAG
- high-dimensionality: $\log(p) \ll n$
- sparsity: maximal degree $d = \max_j |\mathrm{ne}(j)|$ satisfies $d \log(p)/n \to 0$
- "coherence": maximal (partial) correlations $\leq C < 1$
  $\max\{|\rho_{i,j|S}|;\ i \neq j, |S| \leq d\} \leq C < 1$
- signal strength/strong faithfulness:
  $\min\{|\rho_{i,j|S}|; \rho_{i,j|S} \neq 0,\ i \neq j, |S| \leq d\} \gg \sqrt{d \log(p)/n}$

Then, for some suitable tuning param. (level of the tests) and $0 < \delta < 1$:

$$\mathbb{P}[\widehat{\mathrm{CPDAG}} = \text{ true CPDAG}] = 1 - O(\exp(-Cn^{1-\delta}))$$

Sketch of proof

- low-order partial correlations are equivalent to low-dimensional regression parameters
  Gaussian assumption $\rightsquigarrow$ exponential inequality for concentration
- maximal degree of the graph $\rightsquigarrow$ maximal order of partial correlations (maximal dimension of regressions)
- at most $O(\binom{p}{d})$ different partial correlations $\rightsquigarrow$ Bonferroni/union bound with factor $O(d \log(p))$

$\rightsquigarrow$ can show that estimated version of the algorithm "is close" to population version... (some subtle details need to be taken care of)

note that the sample version of the PC-algorithm is order-dependent
$\rightsquigarrow$ "Order-Independent Constraint-Based Causal Structure Learning" (Colombo & Matthuis, 2014)
`https://www.jmlr.org/papers/volume15/colombo14a/colombo14a.pdf`

The role of “sparsity”

as usual: sparsity is necessary for accurate estimation in presence of noise

but here: “sparsity” (so-called protectedness) is crucial for identifiability as well

X causes *Y*

Y causes *X*

cannot tell from observational data the direction of the arrow

the same situation arises with a full graph with more than 2 nodes

$\rightsquigarrow$

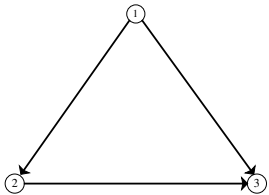
identifiability improves with “sparsity”

Maximum likelihood estimation

without requiring strong faithfulness!

consider Gaussian model $\rightsquigarrow$ Gaussian likelihood

Gaussian P which is Markov w.r.t. DAG D is Gaussian linear structural equation model (see more details later):



$$X_1 \leftarrow \varepsilon_1$$
$$X_2 \leftarrow \beta_{21} X_1 + \varepsilon_2$$
$$X_3 \leftarrow \beta_{31} X_1 + \beta_{32} X_2 + \varepsilon_3$$

$$X_j \leftarrow \sum_{k=1}^{p} \beta_{jk} X_k + \varepsilon_j \ (j = 1, \ldots, p), \ \beta_{jk} \neq 0 \Leftrightarrow \text{ edge } k \to j$$

$$X = BX + \varepsilon, \ \varepsilon \sim \mathcal{N}_p(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)) \text{ in matrix notation}$$

$$X = BX + \varepsilon$$

non-zeroes of B $\Rightarrow$ knowledge of the corresponding DAG

if we would know the order of the variables
$\rightsquigarrow$ (high-dimensional) multivariate regression

but we don't know the order of the variables:

- can only identify equivalence class of B's $\rightarrow$ "`obvious`"
- neg. log-likelihood is non-convex fct.(B) $\rightarrow$ `next slides`
- learning of ordering has large complexity (in general of order p!)

ℓ_0-penalized MLE

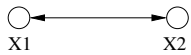
proposed and analyzed for fixed $p < \infty$ by Chickering (2002)

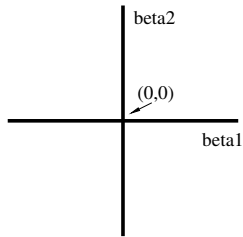
$$\hat{B}, \{\hat{\sigma}_j^2\} = \text{argmin}_{B;\ \{\sigma_j^2\}} - \ell(B, \{\sigma_j^2\};\ \text{data}) + \lambda \underbrace{\|B\|_0}_{\sum_{jk} I(B_{jk} \neq 0)}$$

under the non-convex constraint that B corresponds to "no directed cycles"

Toy-example

$$X_1 \leftarrow \beta_1 X_2 + \varepsilon_1$$
$$X_2 \leftarrow \beta_2 X_1 + \varepsilon_2$$





non-convex parameter space!
(convex relaxation?)

Chickering's (2002) main and important contribution:
algorithm which proceeds greedily on Markov equivalence classes (which is the natural parameter space)

$\rightsquigarrow$ GES (Greedy Equivalent Search)
which in general would not find a global optimum
but Chickering (2002) proves consistency with BIC in low-dimensional problems

Why ℓ_0-penalty?

- ensures the same score for Markov-equivalent structures
  (this would not be true when using ℓ_1-norm penalty)
- ℓ_0-penalty leads to decomposable score

$$\text{score}(D, \mathbf{X}) = \sum_{j=1}^{p} g_j(\mathbf{X}_j, \mathbf{X}_{\text{pa}_D(j)})$$

$\rightsquigarrow$ dynamic programming for computation if $p \approx 20 - 30$
(not easily possible with ℓ_1-norm penalization)
recall that the estimation problem is non-convex...

Statistical properties for ℓ_0-penalized MLE (van de Geer & PB, 2013)

the estimator:
ℓ_0-penalized MLE for the class of minimal I-MAPs
idealized and cannot be computed; it is not the greedy search algorithm (GES)

- no strong faithfulness required for consistency
- under faithfulness: class of minimal I-MAPs = Markov equivalence class
- another "somewhat weaker" permutation beta-min condition is required
- essentially: can only have consistency for the regime $p = o(\sqrt{n/\log(n)})$ with same error variances (see later): $p = o(n/\log(n))$ suffices

the theory is much harder to develop than for the PC-algorithm... in practice, GES is "perhaps a bit better than the PC-algorithm"; see also Nandy, Hauser & Maathuis (2018)

Asymptotic properties: a summary

- PC-algorithm is consistent in high-dimensional regime
  requires a strong faithfulness assumption (necessary)
- GES: greedy equivalent search with ℓ_0-penalized likelihood score function
  consistent for fixed dimension p with BIC penalty

  remarkable since the algorithm does not compute the BIC regularized MLE;
  the consistency is for the greedy search algorithm
  in terms of asymptotics: very rough result
- ℓ_0-penalized MLE:
  consistent in growing-dimensional but restrictive regime $p \ll n$ requiring a permutation beta-min condition (which is weaker than strong faithfulness)

for a long time the ℓ_0-penalized MLE has been computed heuristically
but this has changed in 2024!

INTEGER PROGRAMMING FOR LEARNING DIRECTED ACYCLIC GRAPHS FROM NON-IDENTIFIABLE GAUSSIAN MODELS

TONG XU[1*], ARMEEN TAEB[2*], SIMGE KÜÇÜKYAVUZ[1], AND ALI SHOJAIE[3]

AN ASYMPTOTICALLY OPTIMAL COORDINATE DESCENT ALGORITHM FOR LEARNING BAYESIAN NETWORKS FROM GAUSSIAN MODELS

TONG XU[1], SIMGE KÜÇÜKYAVUZ[1], ALI SHOJAIE[3], AND ARMEEN TAEB[2]

both on arxiv since April and August 2024, respectively

What has been found empirically

- estimating the undirected skeleton of the Markov equivalence class is OK
  the difficulty is the estimation of directionality: and GES (old version) seems empirically a bit better for directionality than PC
- the above point above suggests hybrid algorithms:
  ARGES = Adaptive Restricted Greedy Equivalent Search

  Nandy, Hauser & Maathuis (2018)

  the idea is to restrict GES to a space which is compatible with an initial undirected skeleton of the Markov equivalence class or an undirected conditional independence class (the latter can be estimated by e.g. the nodewise Lasso)

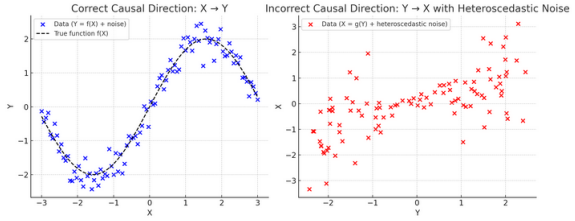
  good empirical performance (like GES)
  consistency in the high-dimensional regime $p \gg n$ under a strong faithfulness assumption

Route via structural equation models: interesting conceptual extensions

full identifiability (card(Markov equivalence class) = 1): if

- same error variances:
  $X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} B_{jk} X_k + \varepsilon_j,\ \mathrm{Var}(\varepsilon_j) \equiv \omega^2$ (Peters & PB, 2014)
- nonlinear structural equation models with additive noise:
  $X_j \leftarrow$ non-linear function $f(X_{\mathrm{pa}(j)}) + \varepsilon_j$
  Mooij, Peters, Janzing & Schölkopf (2009-2012)

That is not a very convincing plot: I would like to see much more heteroscedastic error in the reverse direction. Can you give me such an illustration?



additive noise model: a “more practical” example

Causal Additive Model (CAM)
$X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} f_k(X_k) + \varepsilon_j$ (PB, Ernest & Peters, 2014)

- linear structural eqns. with non-Gaussian errors (LINGAM):
  linear SEM but all $\varepsilon_1, \ldots, \varepsilon_p$ non-Gaussian (Shimizu et al., 2006)

$$X = BX + \varepsilon \iff (I - B)X = \varepsilon$$
$$\rightsquigarrow \quad AX = \varepsilon,\ \varepsilon \text{ independent entries} \implies \text{ICA !}$$

What about hidden variables?

- deconfounding with trim transform is not directly applicable
  because the framework assumes that all X are ancestors of Y
  (upstream of Y)
- work on assuming low-rank structure:

  Frot, Nandy & Maathuis (2019) consider PC with input-covariance matix
  estimated by low-rank constraint (Chandrasekaran et al., 2012)

  direct approach in likelihood scoring by assuming interventional data
  (Taeb, Gamella, Heinze-Deml & PB, 2021)
- various approaches when having interventional data – see later

Open problems and conclusions

open problems:

- elegant and insightful theory for graph recovery and consequences for causal effect estimation
- validation of graph accuracy:
  Hamming distance is too simple-minded
  structural intervention distance (Peters & PB, 2015) is perhaps too complicated
- linear-nonlinear (partially linear) SEMs are complicated in terms of identifiability, and poorly understood(Rothenhäusler, Ernest & PB, 2018)

  with using nonlinear/non-Gaussian SEMs: we bet on additional identifiability – but we should have methods which automatically "adapt" to whether structures are identifiable or not
  ($\rightsquigarrow$ see also later)

conclusions:

- fitting graph equivalence classes from data is hard
- empirically poor performance in comparison to undirected Gaussian graphical models (aka linear model regression)

  insightful theoretical reasons are still missing
  perhaps issues with non-faithfulness or “permutation beta-min condition”
- identifiability is subtle and might has implications on finite sample performance (“near non-identifiability”)
- fully nonlinear and non-Gaussian SEMs lead to perfect identifiability
  interesting trade-off between identifiability and more difficult non-linear estimation

Some selected references:

- Chickering, M. (2002). Optimal structure identification with greedy search. Journal of Machine Learning Research 3, 507-554.
- Hoyer, P., Janzing, D., Mooij, J., Peters, J. & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. Advances in neural information processing systems.
- Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research 8, 613-636.
- Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H. and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47 (11), 1-26.
- Pearl, J. (2000). Causality: Models, Reasoning and Inference. Springer.
- Shimizu, S., Hoyer, P., Hyvärinen, A., Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research 7, 2003-2030.
- Spirtes, P., Glymour, C. and Scheines, R. (2000). Causation, Prediction, and Search. MIT Press.