Sequential Treatments and Causal Discovery

Robin J. Evans and Vanessa Didelez

University of Oxford and BIPS, University of Bremen

APTS Week 4, Oxford
September 2024

Outline

1. Graphical Sequential Treatment
   - Multiple Regression
   - Multiple Treatments

2. Causal Model Selection
   - Markov equivalence
   - Causal Discovery

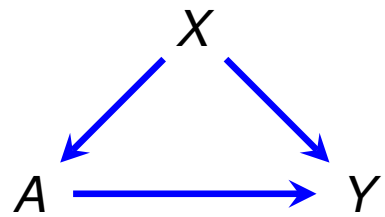
Regression

Question

Suppose we perform a regression of an outcome Y on **several** other variables. Under what circumstances are the coefficients estimating a **causal** quantity?

If they are, then what quantity is it, exactly?

If we compute
$\mathbb{E}[Y \mid A, X] = \beta_{AY \cdot X} A + \beta_{XY \cdot A} X$
under this graph, will we have:



$$p(y \mid do(a)) = \beta_{AY \cdot X} a \, ?$$

$$p(y \mid do(x)) = \beta_{XY \cdot A} x \, ?$$

$\beta_{XY \cdot A}$ **is** a causal effect, but it is the effect of X on Y when keeping A constant; this is called the **controlled direct effect**.

$\beta_{AY \cdot \boldsymbol{C}}$ denotes the regression coefficient for A on Y in the model that includes $\boldsymbol{C}$.

Table 2 Fallacy

JOURNAL ARTICLE

The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients

Daniel Westreich ✉, Sander Greenland Author Notes

American Journal of Epidemiology, Volume 177, Issue 4, 15 February 2013, Pages 292–298,
https://doi.org/10.1093/aje/kws412
Published: 30 January 2013 **Article history** ▾

Abstract

It is common to present multiple adjusted effect estimates from a single model in a single table. For example, a table might show odds ratios for one or more exposures and also for several confounders from a single logistic regression. This can lead to mistaken interpretations of these estimates. We use causal diagrams to display the sources of the problems. Presentation of exposure and confounder effect estimates from a single model may lead to several interpretative difficulties, inviting confusion of direct-effect estimates with total-effect estimates for covariates in the model. These effect estimates may also be confounded even though the effect estimate for the main exposure is not confounded. Interpretation of these effect estimates is further complicated by heterogeneity (variation, modification) of the exposure effect measure across covariate levels. We offer suggestions to limit potential misunderstandings when multiple effect estimates are presented, including precise distinction between total and direct effect measures from a single model, and use of multiple models tailored to yield total-effect estimates for covariates.

Multiple Exposures

Often in applications not a clear distinction between specific exposure and covariates used for adjustment.

There may well be multiple questions of interest: e.g. what is effect of smoking, diet, alcohol intake all **together**.

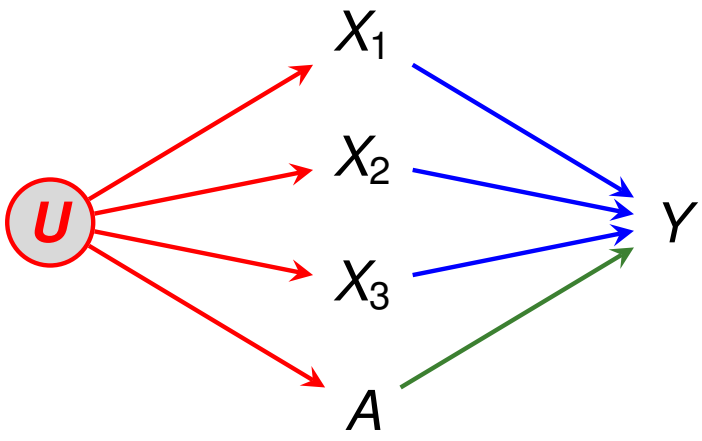
Be clear about causal question relating to multiple exposures: what would be your **ideal target trial**?

Many possible causal effects to define with multiple exposures:

- separate total effects;
- joint intervention effects;
- controlled direct effects;
- strategy for dynamic treatment effects;
- separable (or natural) direct and indirect effects.

Confounding Due to Common History

Suppose the causal structure is more like below—what should we adjust for?



Answer

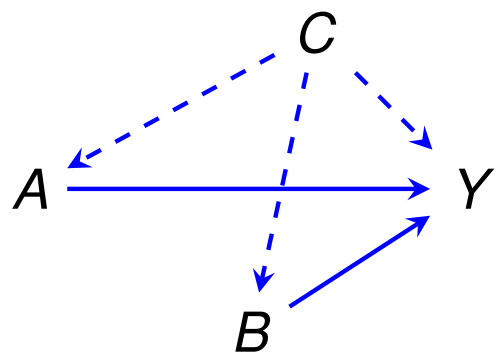
Condition on X_1, X_2, **and** X_3!

This is the only way to block all back-door paths from A to Y.

Examples I

What is the interpretation of β_A, β_B in the regressions?

$$\mathbb{E}[Y \mid A = a, B = b, C = c] = \beta_A a + \beta_B b + \beta_C c.$$

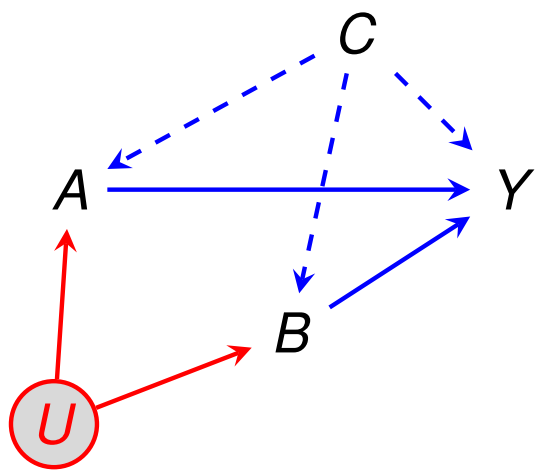


Note that $Y(a, b) \perp_d A, B \mid C$, so indeed (β_A, β_B) **are** the joint causal effects of A, B on Y.

Examples II

What is the interpretation of β_A, β_B in the regressions?

$$\mathbb{E}[Y \mid A = a, B = b, C = c] = \beta_A a + \beta_B b + \beta_C c.$$



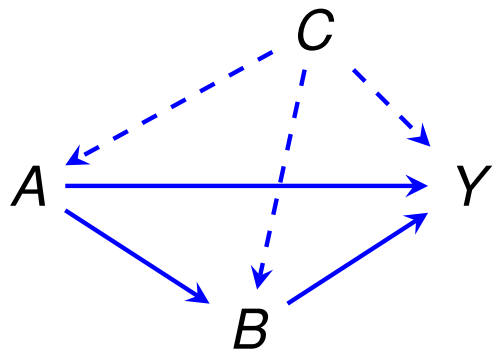
Is $Y(a, b) \perp_d A, B \mid C$?

Yes!

Examples III

What is the interpretation of β_A, β_B in the regressions?

$$\mathbb{E}[Y \mid A = a, B = b, C = c] = \beta_A a + \beta_B b + \beta_C c.$$



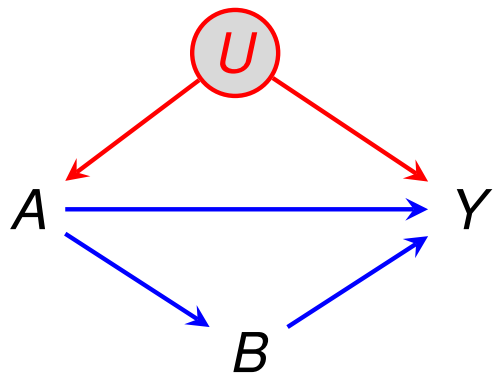
Is $Y(a, b) \perp_d A, B \mid C$?

Yes! **But:** β_A is a **controlled direct effect**, not a total effect.

Examples IV

What is the interpretation of β_A, β_B in the regression:

$$\mathbb{E}[Y \mid A = a, B = b] = \beta_A a + \beta_B b \quad ?$$



Is $Y(a, b) \perp_d A, B$?

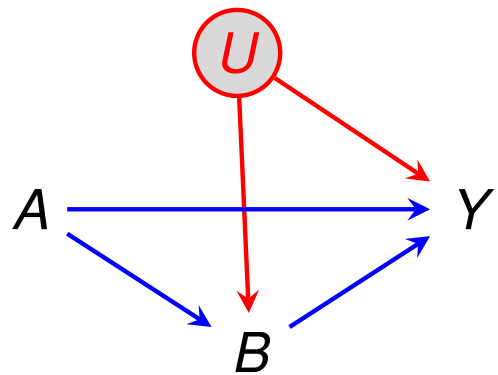
No! β_A has no causal interpretation.

However, $Y(b) \perp_d B \mid A$, so β_B **is** total effect of B on Y.

Examples V

What is the interpretation of β_A, β_B in the regression:

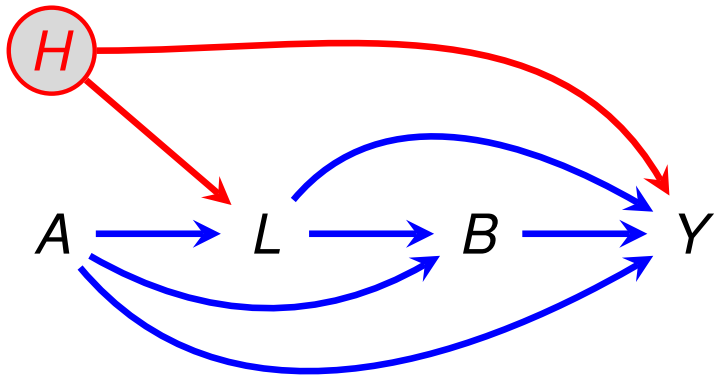
$$\mathbb{E}[Y \mid A = a, B = b] = \beta_A a + \beta_B b \quad ?$$



Is $Y(a, b) \perp_d A, B$?

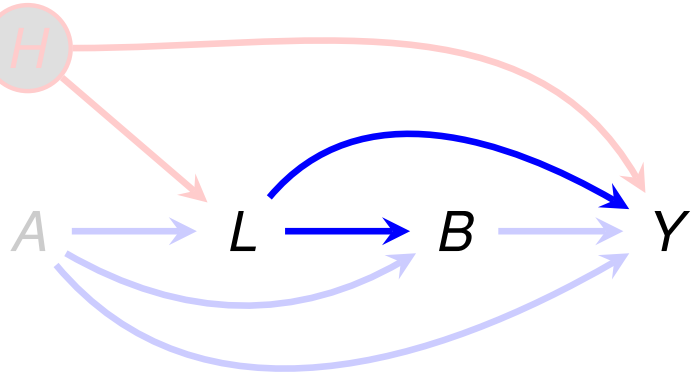
No! Neither coefficient has a causal interpretation.

Sequentially randomized experiment



- A and B are treatments;
- H is unobserved;
- L is a time varying confounder;
- Y is the final response;
- Treatment B is assigned randomly conditional on the observed history, A and L;
- Want to know $P(Y(a, b))$.

Time-Varying Confounding



Should we adjust for L?

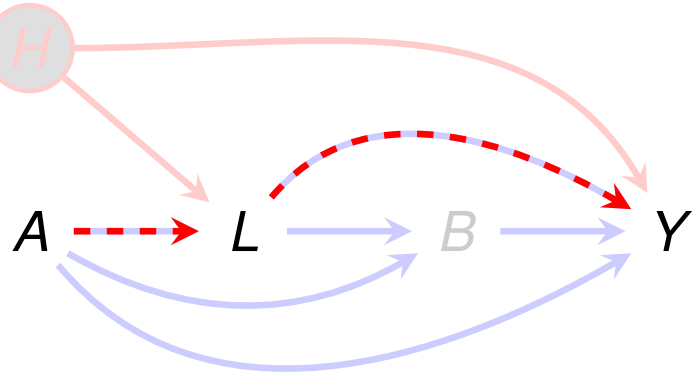
From B's perspective, L needs to be adjusted for to control for confounding.

From A's perspective, L is on the causal path and conditioning would open a spurious path.

Neither regression makes sense!

Need to **break** B's dependence on L in order to estimate the effect.

Time-Varying Confounding



Should we adjust for L?

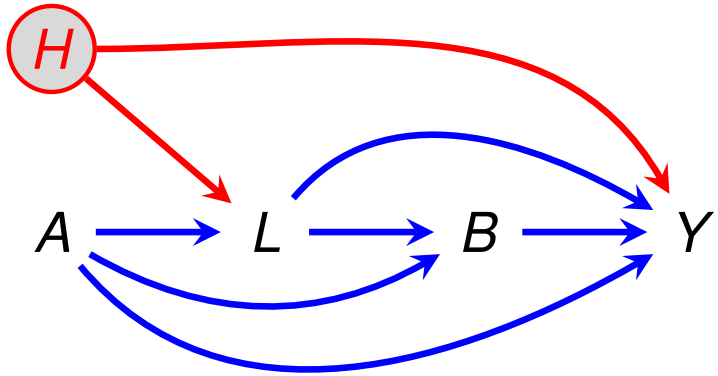
From B's perspective, L needs to be adjusted for to control for confounding.

From A's perspective, L is on the causal path and conditioning would open a spurious path.

Neither regression makes sense!

Need to **break** B's dependence on L in order to estimate the effect.

Sequentially randomized experiment



If the following holds:

$$\begin{aligned} A &\perp\!\!\!\perp Y(a,b) \\ B(a) &\perp\!\!\!\perp Y(a,b) \mid L(a), A \end{aligned}$$

General result of Robins (1986) then implies:

$$P(Y(a,b)=y) = \sum_{\ell} P(L=\ell \mid A=a) \cdot P(Y=y \mid A=a, L=\ell, B=b).$$

Exercise: can you show this?

Do the independences hold?

Marginal Structural Models

Models of the quantity

$$P(Y \mid do(A, B)) = \sum_{\ell} P(L = \ell \mid A) \cdot P(Y \mid A, L = \ell, B)$$

are called **marginal structural models** (Robins et al., 2000).

They have various nice properties:

- can be modelled semi-parametrically (so no need to fully specify rest of the distribution)
- either the propensity scores ($P(A)$ and $P(B \mid A, L)$) **or** the outcome models $P(Y \mid A, L, B)$ and $P(L \mid A)$ can be used to identify parameters.
- there is also a doubly robust version!

Hard to simulate from, though Evans and Didelez (2024) provide a solution!

IPW Identification

The model suffers from time-dependent confounding, so weights must reflect this.

Idea of IPW is to obtain the g-formula by 'removing' pieces not in it. So:

$$P(A, L, B, Y) = P(A) \cdot P(L \mid A) \cdot P(B \mid A, L) \cdot P(Y \mid A, L, B)$$

So create a **pseudo-population** by fitting models for $P(A)$ and $P(B \mid A, L)$ and then reweight ith observation by

$$w_i = \frac{1}{\widehat{P}(A = a_i)} \cdot \frac{1}{\widehat{P}(B = b_i \mid A = a_i, L = \ell_i)}$$

Common to use logistic regression if A, B are binary.

IPW Remarks

Conditions are similar to single treatment case; see Robins et al. (2000) for details.

General form of weights with treatments $A_0, \ldots, A_K$ and covariates $L_0, \ldots, L_K$ is

$$\widehat{w}_i = \prod_{k=0}^{K} \frac{1}{\widehat{P}(A_k \mid \overline{A}_{k-1}, \overline{L}_k)},$$

where $\overline{A}_j = (A_0, \ldots, A_j)$.

- Models for $A_k | \overline{A}_{k-1}, \overline{L}_k$ must be correctly specified;
- in practice people use stabilised weights;
- can also estimate optimal dynamic treatments using *Q-learning* (Chakraborty and Moodie, 2013).

References

Chakraborty, B. and Moodie, E.M. Statistical Methods for Dynamic Treatment Regimes. *Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer, 2013.

Evans, R.J. and Didelez, V. Parameterizing and simulating from causal models (with discussion). *JRSS-B*, 86(3) pp 535–568, 2024.

Robins, J.M., Hernan, M.A. and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp.550-560, 2000.

Outline

1. Graphical Sequential Treatment
   - Multiple Regression
   - Multiple Treatments

2. Causal Model Selection
   - Markov equivalence
   - Causal Discovery

Markov equivalence

Notice that the Markov properties of **distinct** graphs are sometimes the same.

$$p(x) \cdot p(z \mid x) \cdot p(y \mid z)$$

$$p(y) \cdot p(z \mid y) \cdot p(x \mid z)$$

$$p(z) \cdot p(x \mid z) \cdot p(y \mid z)$$

All these graphs imply (precisely) that $X \perp\!\!\!\perp Y \mid Z$.

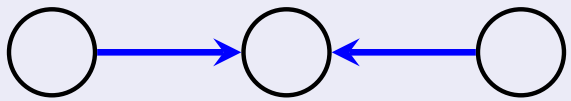
If two different graphs define the same model, we say that they are **Markov equivalent**.

A set of graphs that are all Markov equivalent is called the **Markov equivalence class**.

PDAGs and CPDAGs

This means that, given data, we cannot tell what the precise underlying DAG that generated the data was!

The Markov equivalence class of a DAG is determined by its skeleton (i.e. which nodes are adjacent) and its **unshielded colliders** or **v-structures**; that is induced subgraphs isomorphic to:



Instead we can represent the **pattern**, by orienting any triples with precisely two edges if they are a v-structure.

If the graph is compatible with a DAG we call it a **partially directed acyclic graph** (PDAG).

Other orientations follow logically by Meek's rules (see Appendix). Applying exhaustively means we obtain a **completed** PDAG (CPDAG).

What can we do if we don't know the model?

Typically, we assume that there **is** some DAG model that explains the data we observe.

There are then various approaches we can use:

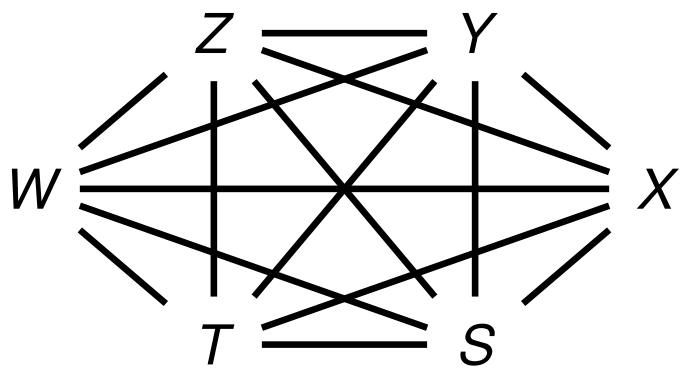
- learn the DAG from observational data alone
  (e.g. PC/FCI algorithm, GES, order/partition MCMC, SAT solvers);
- if available, use data from interventions or perturbations of the model (e.g. invariant prediction, GIES);
- use non-linearities or non-Gaussian errors to exploit asymmetries
  (e.g. LiNGAM, causal additive noise models).

Constraint-Based Approaches

These use signficance tests to decide whether constraints hold or not.

The **PC algorithm** (Spirtes et al. 2000) works by starting with a complete graph, and testing all marginal independences.

Any hypothesis **not** rejected (at size α) leads to the removal of an edge, and the empty conditioning set is recorded.



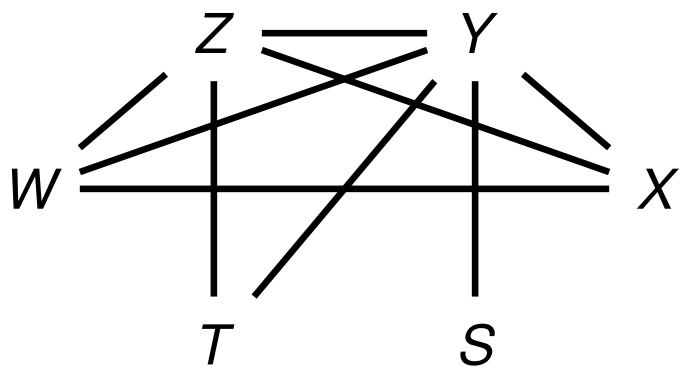
The algorithm then proceeds to conditioning sets of size 1, 2, etc., searching only through the neighbours of one of the two vertices.

Constraint-Based Approaches

These use signficance tests to decide whether constraints hold or not.

The **PC algorithm** (Spirtes et al. 2000) works by starting with a complete graph, and testing all marginal independences.

Any hypothesis **not** rejected (at size α) leads to the removal of an edge, and the empty conditioning set is recorded.



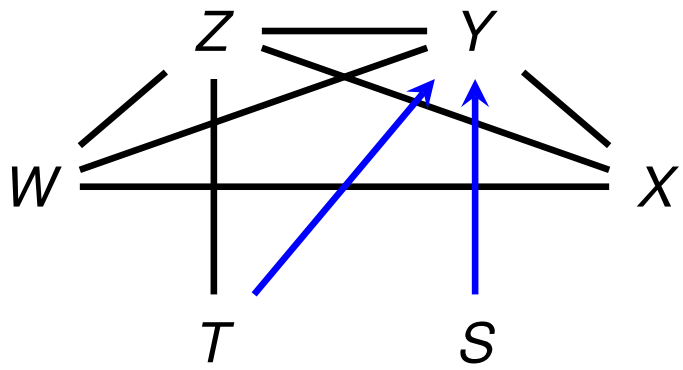
The algorithm then proceeds to conditioning sets of size 1, 2, etc., searching only through the neighbours of one of the two vertices.

Constraint-Based Approaches

These use signficance tests to decide whether constraints hold or not.

The **PC algorithm** (Spirtes et al. 2000) works by starting with a complete graph, and testing all marginal independences.

Any hypothesis **not** rejected (at size α) leads to the removal of an edge, and the empty conditioning set is recorded.



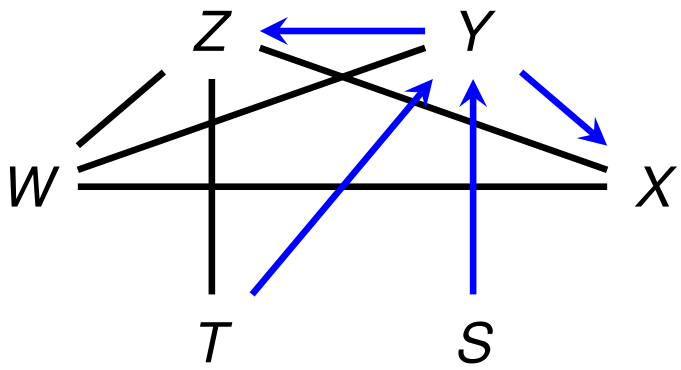
After determining the skeleton, it goes back to see how unshielded triples should be oriented based on the set that made the variables conditionally independent.

Constraint-Based Approaches

These use signficance tests to decide whether constraints hold or not.

The **PC algorithm** (Spirtes et al. 2000) works by starting with a complete graph, and testing all marginal independences.

Any hypothesis **not** rejected (at size α) leads to the removal of an edge, and the empty conditioning set is recorded.



After determining the skeleton, it goes back to see how unshielded triples should be oriented based on the set that made the variables conditionally independent.

It then uses Meek's rules to obtain the CPDAG.

Constraint-Based Approaches

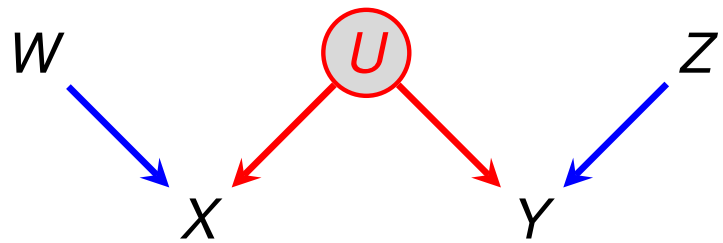
This algorithm is **very efficient** if the graph is sparse.

The PC algorithm requires an assumption of **faithfulness** to work. That is:

$$X \perp\!\!\!\perp Y \mid Z\,[p] \qquad \Rightarrow \qquad X \perp_d Y \mid Z\,[\mathcal{G}].$$

This is the converse of p being **Markov** w.r.t. $\mathcal{G}$.

Since $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$ can look very similar (Evans, 2020), it will typically remove edges too freely, and it is **not guaranteed** to return a CPDAG!



Alternatives (e.g. conservative PC) also exist.

Score-Based Approaches

Score-based approaches attempt to maximize a score that will be consistent for the true model.

Example

The classic score-based approach for DAGs is **greedy equivalence search** (GES).

This has two phases; starting with an empty graph:

1. add the edge to the PDAG that maximizes the local increase in the score, stopping when no improvement is possible;
2. remove the edge that maximizes the local increase in the score, stopping when no improvement is possible.

For sufficiently large samples, GES is guaranteed to find the Markov equivalence class of the true DAG (Chickering, 2002).

Hybrid Approaches

PC and GES both have high-dimensional guarantees (e.g. Kalisch and Bühlmann, 2007, Nandy et al., 2018).

However, we generally have the following trade-off to make:

- constraint-based approaches are fast, but often make mistakes that cascade into further errors;
- score-based approaches are more accurate, but take longer to run.

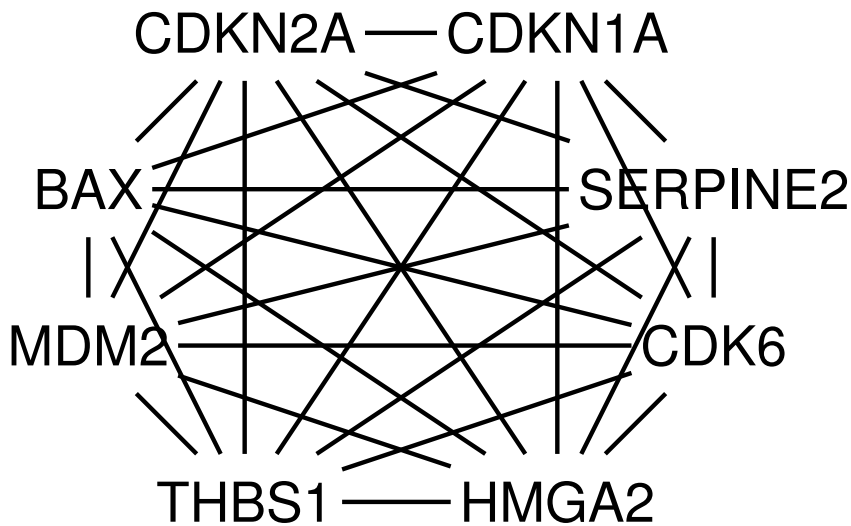
This motivates **hybrid approaches**, which typically start by reducing the search space using constraints, and then proceed to use a more accurate score-based method on the smaller set of models.

Examples include Max-Min Hill Climbing, or running the first phase of PC and then GES on the resulting skeleton.

See Scutari and Denis (2021) for examples in R.

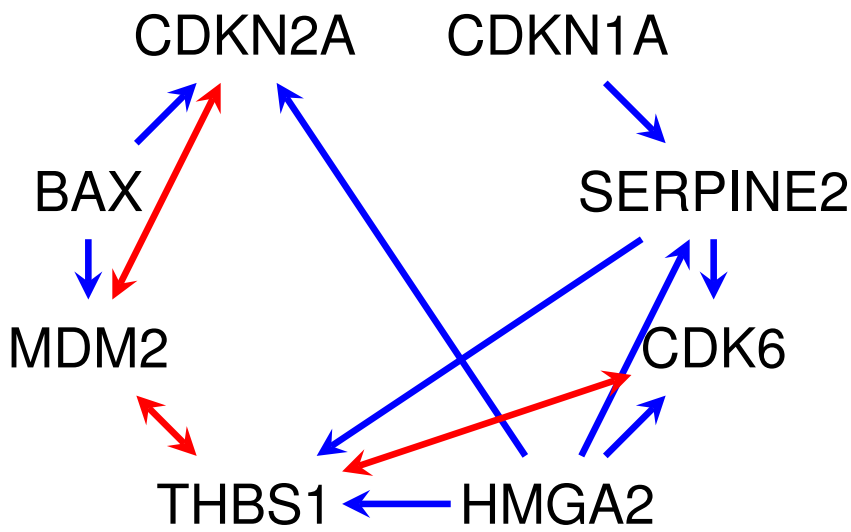
Causal Search Example

The dataset `tcgas` contains expression data from eight genes in tumour tissues taken from $n = 362$ patients with squamous cancer of the head and neck (Foraita et al., 2020).



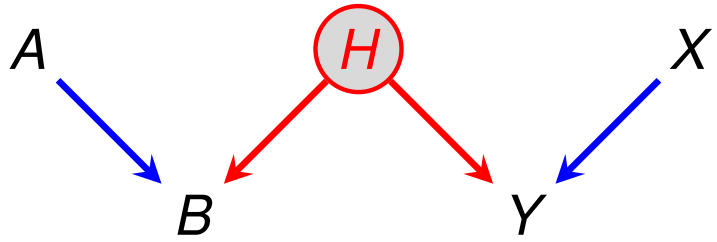
Causal Search Example

The dataset `tcgas` contains expression data from eight genes in tumour tissues taken from $n = 362$ patients with squamous cancer of the head and neck (Foraita et al., 2020).



Hidden Variables

There are configurations of independences from DAGs that give these **bidirected** edges, but they require **unobserved variables**.



This extends the ‘projection’ operation given in the efficient adjustment section.

The result on the previous slide suggests that hidden variables should be allowed for, which would mean using the **fast causal inference** (FCI) algorithm instead (Spirtes et al., 2000).

LiNGAM

All the approaches so far can only discover the **Markov equivalence class** of the model.

Other assumptions and new methods are required to determine causal directions within these classes.

One such approach is **Linear, Non-Gaussian Acyclic Models** (LiNGAM), which assumes that the noise is **not** Gaussian.

The idea is then very similar to **independent components analysis** (ICA), where the sum of non-Gaussian variables can be disentangled.

We then have, for $X \in \mathbb{R}^p$,

$$X = BX + \varepsilon \quad \Longrightarrow \quad X = (I - B)^{-1}\varepsilon.$$

ICA aims to deduce $W = I - B$, and hence obtain the original sources $\varepsilon \in \mathbb{R}^p$.

LiNGAM Algorithm

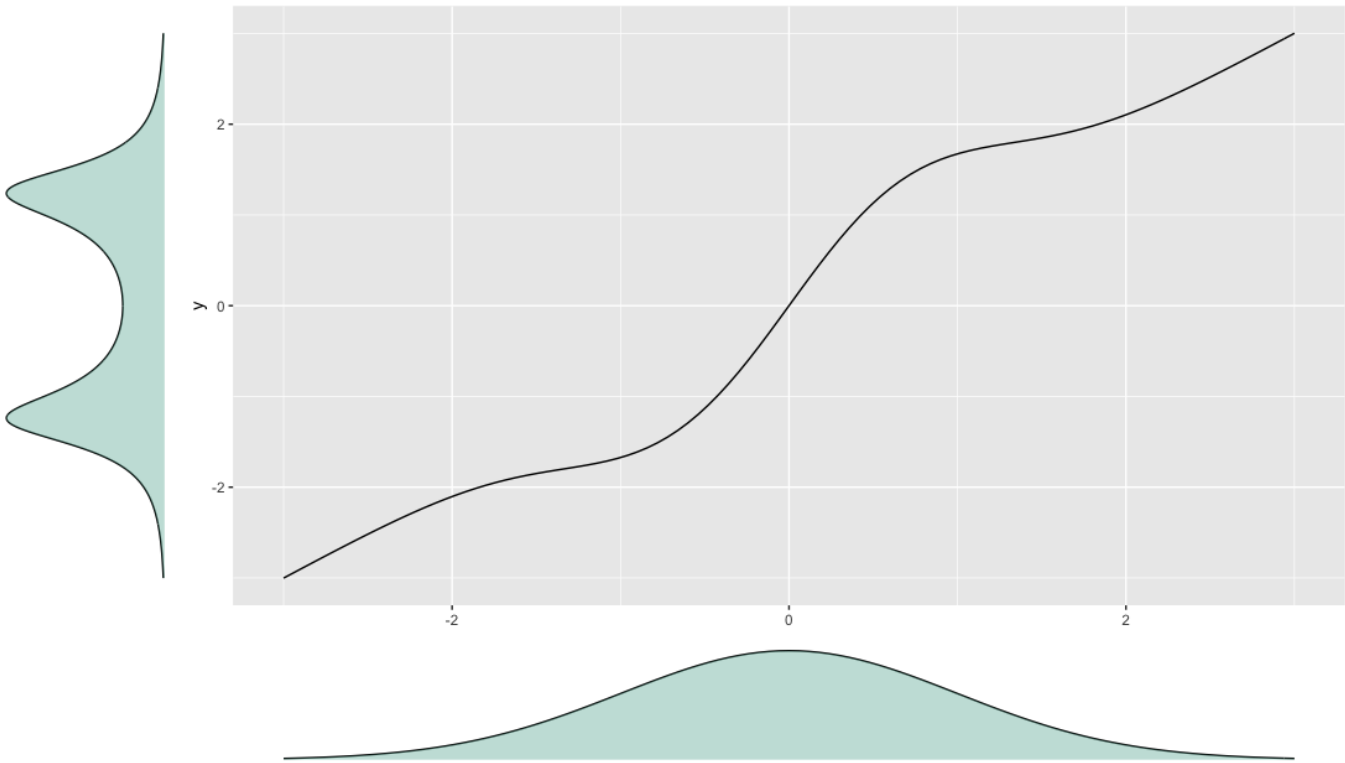
In ICA, we are given n samples of a p-dimensional vector and asked to 'disentangle' the sources.

It is crucial that none (or at most one) of the sources is Gaussian, otherwise the original sources become unidentifiable.

1. Given an $n \times p$ data matrix X ($n \gg p$), centre each column.
2. Apply the ICA algorithm $X = SW^{-1}$ where $S \in \mathbb{R}^{n \times p}$ contains independent components in its rows.
3. Find unique permutation of columns of W s.t. no zeros on diagonal.
4. Divide each column of W by its corresponding diagonal element.
5. Compute an estimate $\hat{B} = I - \tilde{W}$.
6. Get permutation of $\hat{B}$ that is lower triangular.

Information Geometric Approach

For this approach we need a **postulate** (assumption): that the **mechanisms** that generate values of the **cause** should be unrelated to the mechanism that turns causes into **effects** (Daniusis, 2010).

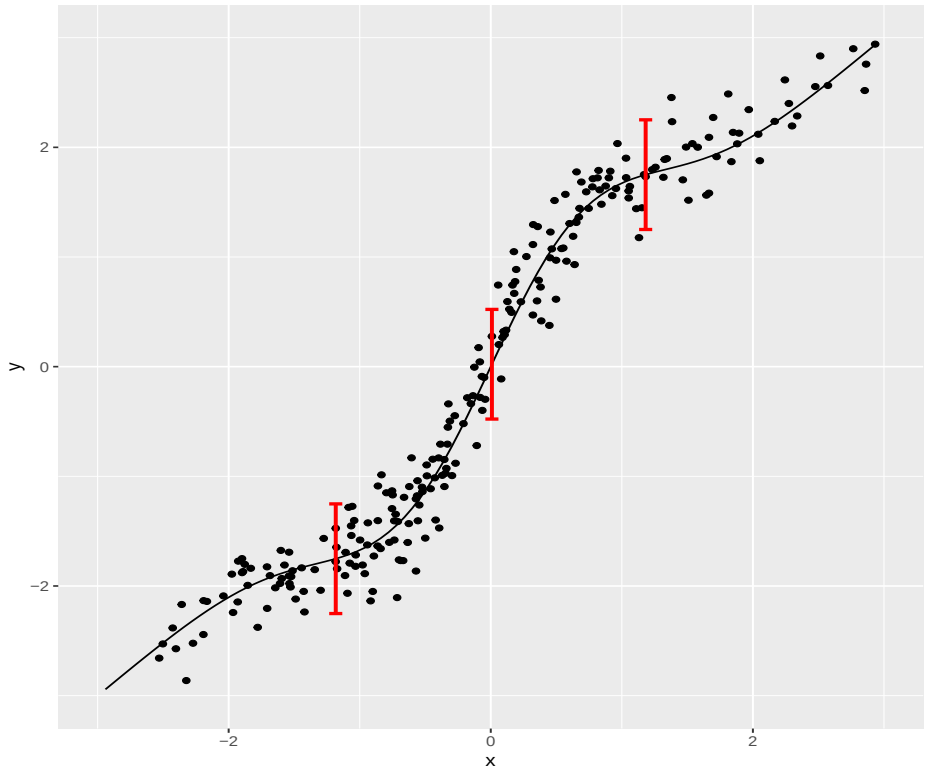


It will then follow that the distribution of the cause is unrelated to the mechanism, while the effect **is** 'correlated' with it.

Additive Noise Models

Another approach is to assume an **additive noise model**:

$$Y = f(X) + \epsilon, \qquad \epsilon \perp\!\!\!\perp X.$$

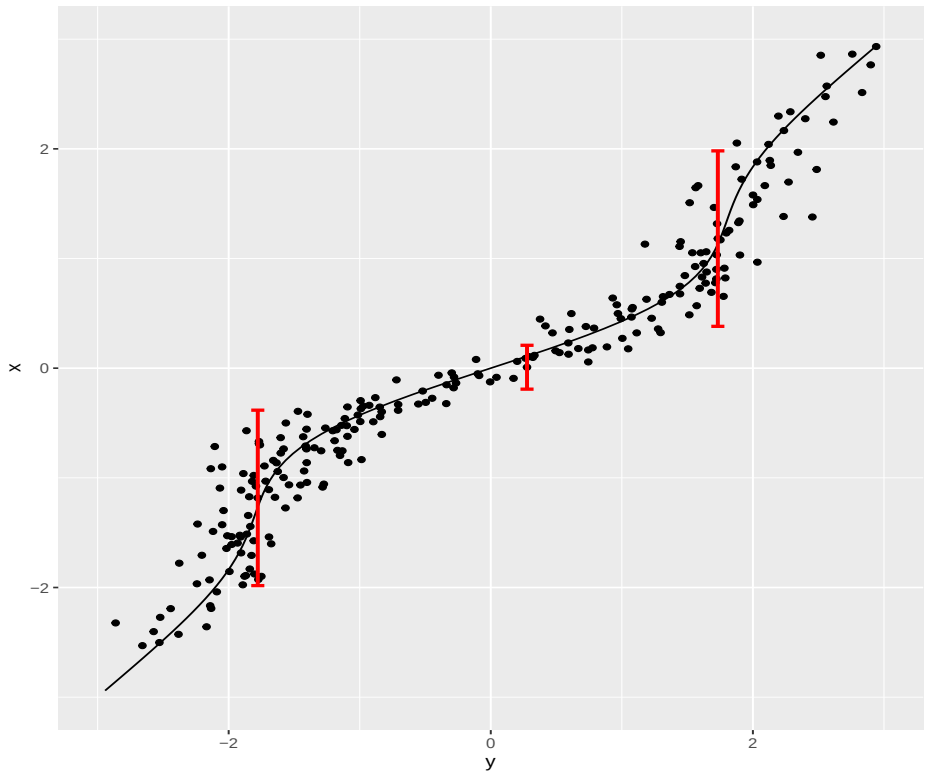


This way around the noise distribution looks similar at all points.

Additive Noise Models

However, if we flip the picture...

$$X \neq g(Y) + \delta, \qquad \delta \perp\!\!\!\perp Y.$$



...now clear dependency between steepness of the curve and variance.

RESIT

This motivates the Regression and Subsequent Independent Test (RESIT) algorithm of Peters et al. (2014).

First, perform a (flexible) regression of Y on X to give residuals r_Y and fitted values f_X.

Then do the same for X on Y to give residuals r_X and fitted values f_Y.

Then see whether $r_Y \perp\!\!\!\perp f_X$ or $r_X \perp\!\!\!\perp f_Y$ is more plausible, and determine the causal direction accordingly.

Note that, for this method to work, we **cannot** have a linear function $f : X \mapsto Y$ **and** a Gaussian noise distribution (the only symmetric case!)

NO TEARS

The optimization procedure we have described is effectively **discrete**, because we have to search over a space of graphs.

Several recent methods (Zheng et al., 2018) instead use a **continuous function** to enforce acyclicity. For example:

$$\text{tr}(\exp\{A \circ A\}) - d = 0 \quad \iff \quad \mathcal{G} \text{ is acyclic,}$$

where A is the $d \times d$-adjaceny matrix for $\mathcal{G}$, and $\circ$ represents pointwise multiplication.

Combined with an assumption of additive noise models makes NO TEARS a state-of-the-art approach.

Other approaches **combine** a structure learning method (such as NO TEARS) with methods for estimating causal effects. See, for example, Geffner et al. (2022) or Sharma and Kiciman (2020).

Criticized by Reisach et al. (2021) as just checking '**varsortability**'.

Other Methods

There are **countless** other methods that exist (Vowels et al., 2022):

Method	Year	Type	Suff.	Faith.	Acycl.	Interv.	Output
PC [228]	1993	constraint	yes	yes	yes	no	CPDAG
FCI [2							
CCD [							
TPDA							
CPC [							
KCL [2							
ION [2							
IDA [1							
GSF [103]							
bQCD [235]							
GCL [251]							
GGIM [55]							
GYKZ [68]							
SLARAC etc. [260]							
Order-MCMC [60]							
OG [53]							
EE-DAG [281]							
ZIPBN [35]							
LiNGAM [221]							
LV LiNGAM [101]							
non-linear ANM [							
PNL [277]							
CAN [113]							
CCM [232]							
IGCI [112]							
KCDC [161]							
MMHC [245]							
ARGES [172]							

Method	Year	Data	Form	Acycl.	Interv.	Output
EG [52]						
TWILP [18						
CAM [26]						
K2 [38]						
LB-MDL [						
HGC [91]						
CMS [152]	2014	low, dynamic/time series	NN	–	no	direction
NO TEARS [279]	2018	low	linear	yes	no	DAG
CGNN [74]	2018	low	NN	yes	no	DAG
SAM [121]	2019	low/medium	NN	yes	no	DAG
DAG-GNN [272]	2019	low	NN	yes	no	DAG
GAE [178]	2019	low	NN	yes	no	DAG
NO BEARS [142]	2019	low/medium/high	3-poly	yes	no	DAG
DEAR [219]	2020	image	NN	yes	no	–
CAN [167]	2020	low/medium/image	NN	yes	no	DAG
NO FEARS [259]	2020	low	linear	yes	no	DAG
GOLEM [177]	2020	low	linear	yes	no	DAG
ABIC [19]	2020	low	linear	yes	no	ADMG/PAG
DYNOTEARS [179]	2020	low	linear	yes	no	SVAR
SDI [124]	2020	low	NN	yes	yes	DAG
AEQ [63]	2020	Bi	NN	–	no	direction
RL-BIC [284]	2020	low	NN	yes	no	DAG
CRN [125]	2020	low	NN	yes	yes	DAG
ACD [151]	2020	low, time series	NN	Granger	no	time-series DAG
CASTLE (reg.) [138]	2020	low/medium	NN	yes	no	DAG
GranDAG [139]	2020	low	NN	yes	no	DAG
MaskedNN [176]	2020	low	NN	yes	no	DAG
CausalVAE [267]	2020	image	NN	yes	yes	DAG
CAREFL [126]	2020	low	NN	yes	no	DAG / direction
Varando [251]	2020	low	linear	yes	no	DAG
NO TEARS+ [280]	2020	low	non-linear	yes	no	DAG
ICL [258]	2020	low	NN	yes	no	DAG
LEAST [283]	2020	low/medium/high	linear	yes	no	DAG
CausalMosaic [263]	2020	Bi	NN	–	no	direction
NSM [253]	2021	video, dynamic/time series	NN	–	no	direction

References

Chickering, D.M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3, 507-554, 2002.

Evans, R.J. Model selection and local geometry. *Annals of Statistics*, 2020.

Foraita, R. et al. Causal discovery of gene regulation with incomplete data. *JRSS-A*, 183(4), pp.1747-1775, 2020.

Geffner, T. et al. Deep end-to-end causal inference. *arXiv:2202.02195*, 2022.

Kalisch, M. and Bühlman, P. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Nandy, P., Hauser, A. and Maathuis, M. H. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46(6A), 3151-3183, 2018.

Reisach, A., Seiler, C. and Weichwald, S. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. *NeurIPS*, 34, pp.27772-27784, 2021.

References (cont.)

Peters, J., Mooij, J.M., Janzing, D. and Schölkopf, B. Causal discovery with continuous additive noise models. *JMLR* 15:2009–2053, 2014.

Scutari, M. and Denis, J.B. *Bayesian networks: with examples in R*, Chapman and Hall/CRC Press, 2021.

Sharma, A. and Kiciman, E. DoWhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*. 2020.

Spirtes, P., Glymour, C.N. and Scheines, R.. *Causation, Prediction, and Search*. MIT press (2nd edition). 2000

Vowels, M. J., Camgoz, N. C. and Bowden, R. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4), 1-36, 2022.

Zheng, X., Aragam, B., Ravikumar, P.K. and Xing, E.P. DAGSs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31. 2018.