Causal graphical models for galaxy surveys

Serafina Di Gioia
PostDoc in Applied ML @ ICTP

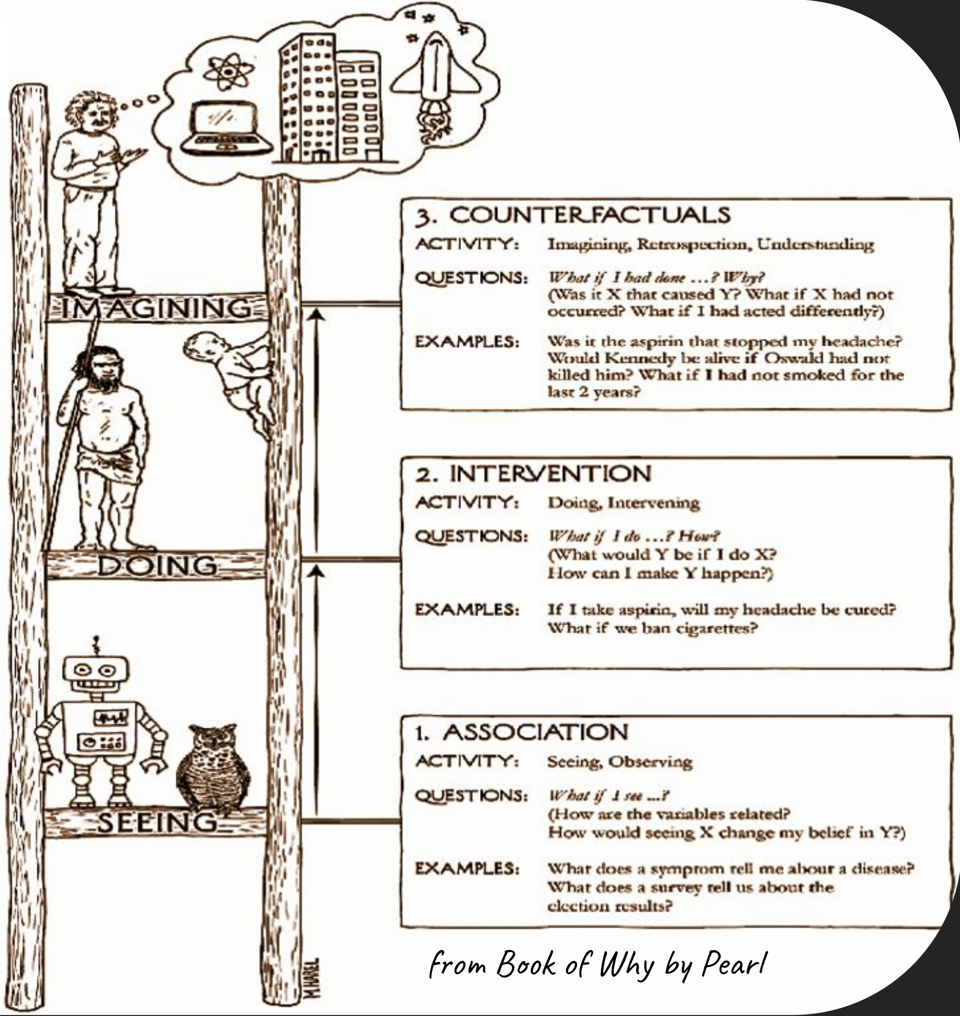
30th November 2023

What is causality?

"Actual" Causality

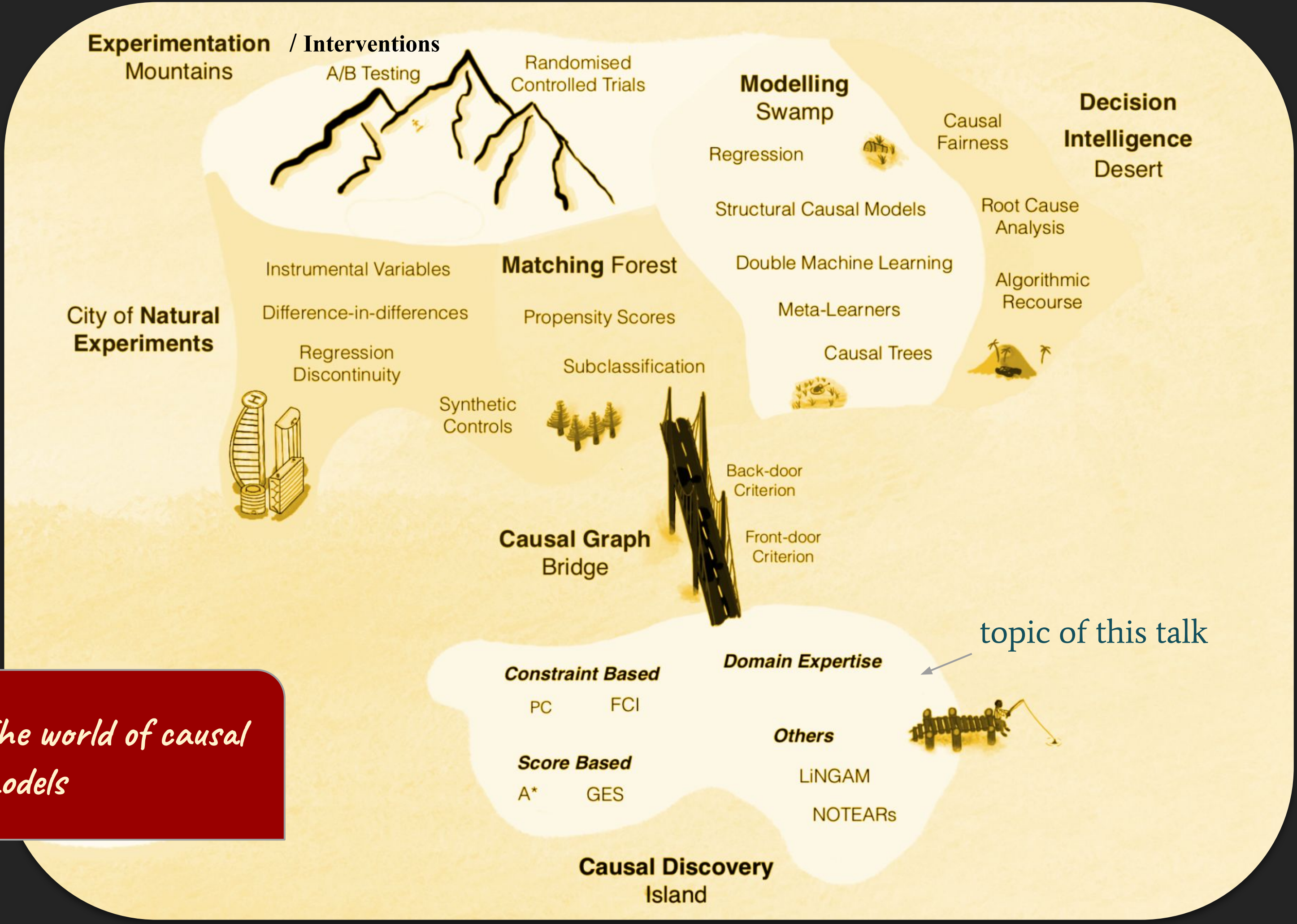
"Causality-in-mean"

Statistics



from Book of Why by Pearl

The world of causal models



Tools for causal discovery

- **THEORY**
  Probabilistic graphical models + causal inference

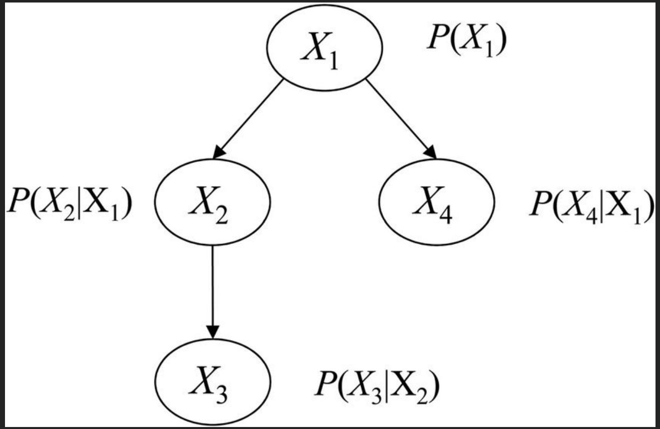
- **COMPUTATIONAL METHODS:**
  causal structure/parameter learning algorithms

Major ingredient of PGMs: Bayesian networks (BN)

Bayesian networks have 2 components:

- directed acyclic graph (DAG)
- Joint probability distribution

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{N} \mathrm{P}(X_i \mid \Pi_{X_i}; \Theta_{X_i})$$



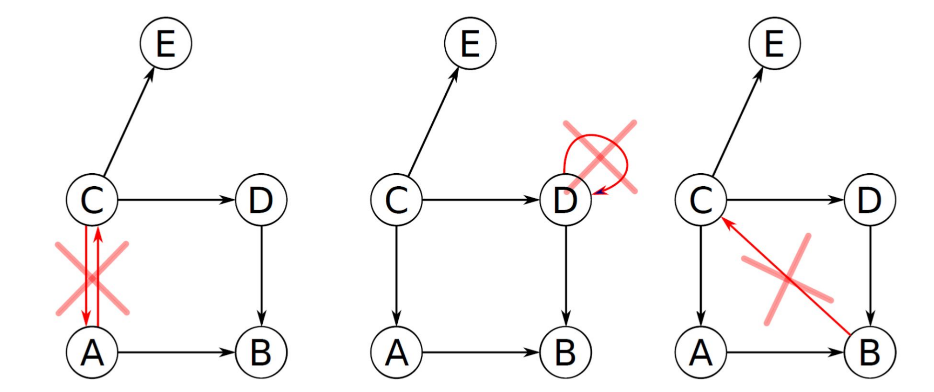
The world of DAGs

We can indicate a graph with the tuple

$$\mathcal{G} = (\mathbf{V}, A)$$

$V = \{v_1 ... v_N\}$ set of nodes

$a_{ij} = (v_i, v_j)$ set of arcs



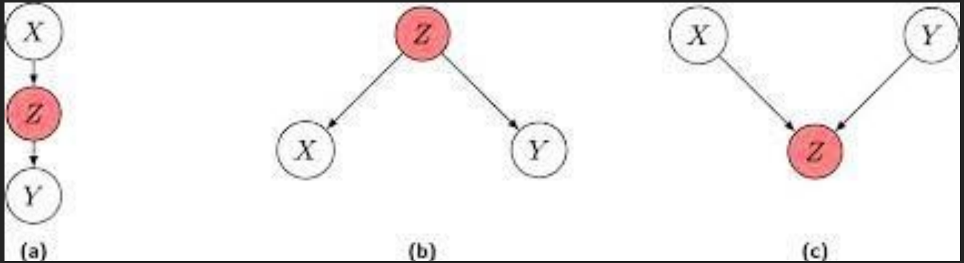
If a graph satisfies:

1. no undirected edges
2. no loop
3. no cycle

is called a Directed Acyclic graph (DAG)

Building blocks of causal BNs

Triplets



chain

fork

collider /
v-structure

$X \perp\!\!\!\perp Y \mid Z$

$X \not\!\perp\!\!\!\perp Y$

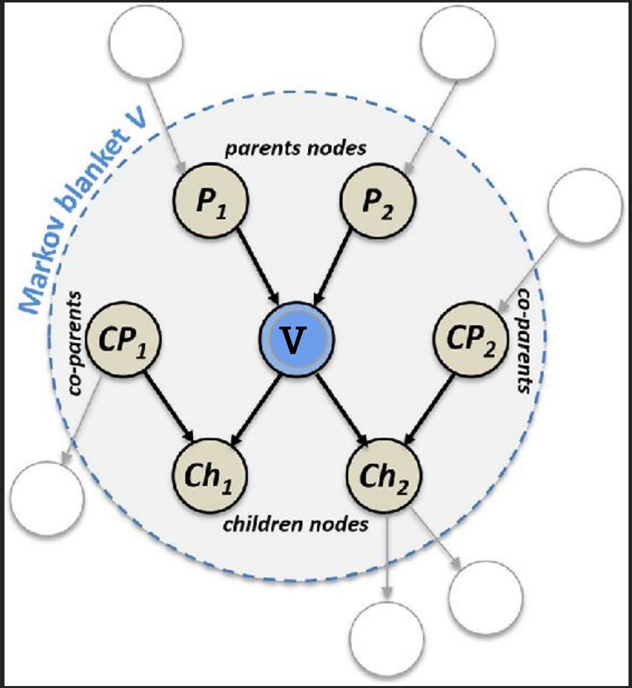
$X \perp\!\!\!\perp Y$

$X \not\!\perp\!\!\!\perp Y \mid Z$

P(X,Y,Z) = P(Y|Z)P(Z|X)P(X)
=P(Y|Z) P(X|Z)P(Z)

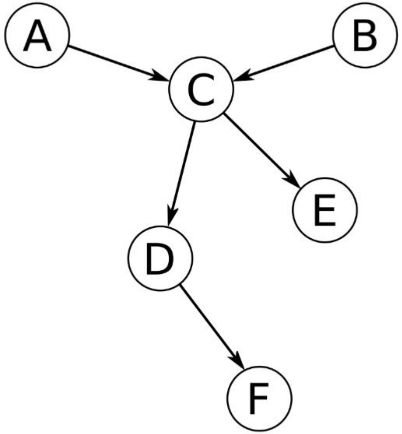
P(X,Y,Z) = P(Z|X,Y) P(X) P(Y)

Markov blanket



Conditional independence is encoded in graph properties

DAG



$\Longrightarrow$

Graphical separation

$A \perp\!\!\!\perp_G B$
$A \perp\!\!\!\perp_G D \mid C$
$B \perp\!\!\!\perp_G D \mid C$
$A \perp\!\!\!\perp_G E \mid C$
$B \perp\!\!\!\perp_G E \mid C$
$D \perp\!\!\!\perp_G E \mid C$
$C \perp\!\!\!\perp_G F \mid D$
$\dots$

$\Longrightarrow$

Probabilistic independence

$A \perp\!\!\!\perp_P B$
$A \perp\!\!\!\perp_P D \mid C$
$B \perp\!\!\!\perp_P D \mid C$
$A \perp\!\!\!\perp_P E \mid C$
$B \perp\!\!\!\perp_P E \mid C$
$D \perp\!\!\!\perp_P E \mid C$
$C \perp\!\!\!\perp_P F \mid D$
$\dots$

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Longleftarrow \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}$$

Causal Structure-learning algorithms:

There are three main classes of algorithms for causal discovery:

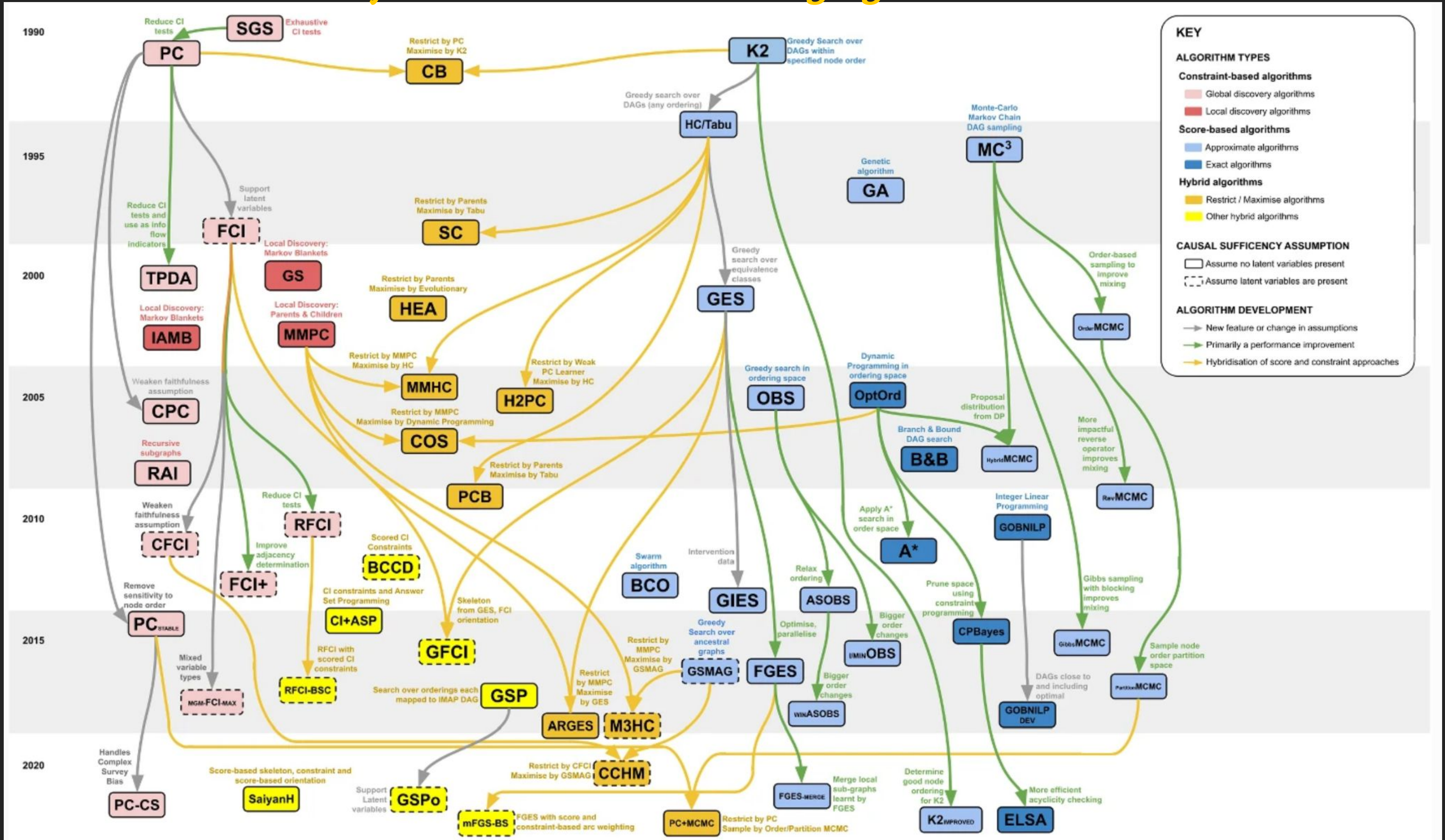
- constraint-based $\xrightarrow{\text{based on}}$ Conditional Independence tests
- score-based $\longrightarrow$ score optimization + grid-search
- hybrid $\longrightarrow$ mixing score optim. with CI tests

General assumptions:

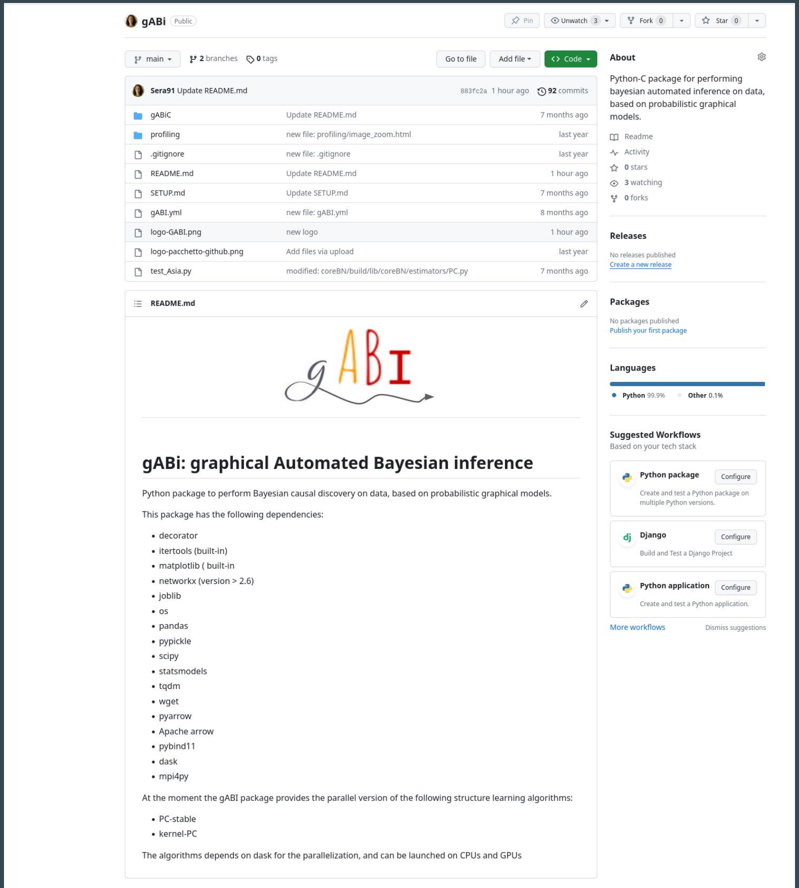
1. Causal Markov
2. Causal sufficiency
3. Causal faithfulness

$\longrightarrow$ { no collinear variables
no missing confounders }

History of causal structure learning algorithms



The gABi library



The gABI library provides three constraint-based algorithms for causal structure learning:

- PC
- kernel-PC
- IAMB

Comparison between PC-stable and kernel-PC

PC-stable is the most common constraint-based algorithm and kernel-PC is its generalization to non-linear ANMs

	Kernel-PC	PC-stable
Differences in assumptions	non-linear relationship gaussian/non-gaussian noise	linear relationships gaussian noise gaussian data
differences in implementation	generalized transitive phase	transitive phase
Conditional Independence test	HSIC-gamma /DCC-gamma HSIC-perm / DCC-perm	Pearson CI test Fisher-Z

Digression on CI tests

For LINEAR GAUSSIAN models described by

$$X_i = \beta_j X_j + \epsilon_i$$

we use PARAMETRIC CI tests (e.g. Fisher-Z test, Exact-t test)

For NON-LINEAR ADDITIVE NOISE models described by

$$X_j = f_j(\mathrm{Pa}_j) + N_j$$

we use kernel NON-PARAMETRIC tests (e.g. HSIC-perm, HSIC-gamma, DCC-gamma)

but in practice...

- **estimating conditional dependence for continuous domain is not straightforward**
- **many algorithms use linear statistical methods or discretization**

Why kernel methods for testing CI?

<u>basic idea</u>

covariance structure on RKHS gives info on dependence and conditional independence of original variables

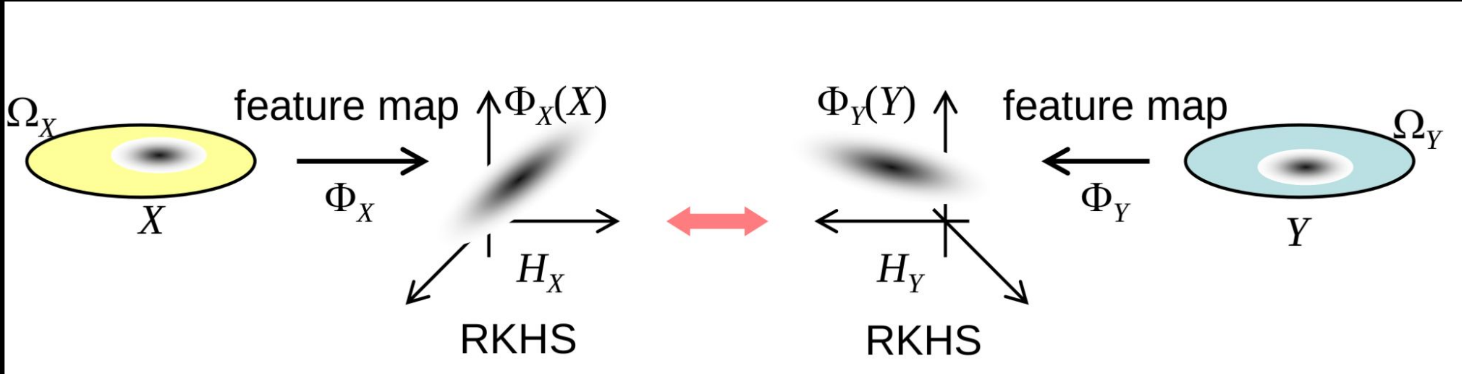
$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathrm{Cov}[f(X), g(Y)]$$

can be decomposed

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$$

$$V_{YX|Z} = V_{YX} - V_{YZ} V_{ZX}$$

$$V_{YX|Z} = \Sigma_{YY}^{-1/2} (\Sigma_{YX} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX}) \Sigma_{XX}^{-1/2}$$



Kernel CI tests: residuals gamma test

Test hypothesis:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y | Z \ \textbf{v.s.} \ \mathcal{H}_1 : X \not\perp\!\!\!\perp Y | Z$$

$$r_X \perp\!\!\!\perp r_Y$$

residuals are estimated with a GAM fit (pyGAM) $y = f_0 + \sum_{i=1}^{p} f_i(x_i) + \varepsilon$

Gamma approximation

$$\mathrm{IC}(x, y) \sim \mathrm{Gam}(\alpha, \theta)$$ where $$\alpha = \frac{\mathrm{E}[\hat{\mathrm{IC}}_{X,Y}]^2}{\mathrm{Var}(\hat{\mathrm{IC}}_{X,Y})}, \quad \theta = \frac{\mathrm{Var}(\hat{\mathrm{IC}}_{X,Y})}{\mathrm{E}[\hat{\mathrm{IC}}_{X,Y}]}$$

p-value: upper-tail quantile of Gam

$$\mathbb{H}(P, Q) = \frac{1}{(N-1)^2} \mathrm{tr}(K_P H K_Q H)$$ where $$K_{i,j} = \exp \frac{(x_i - x_j)^2}{\sigma^2}, \ H_{i,j} = \delta_{i,j} - \frac{1}{n}$$

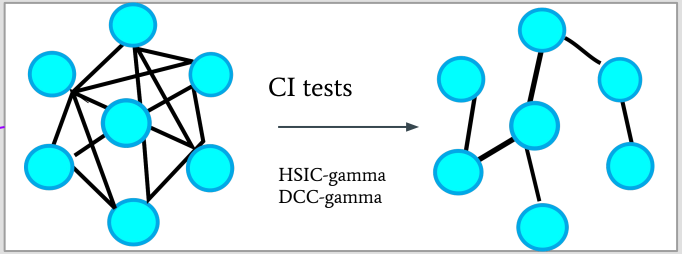
Phases of the kernel-pc algorithm

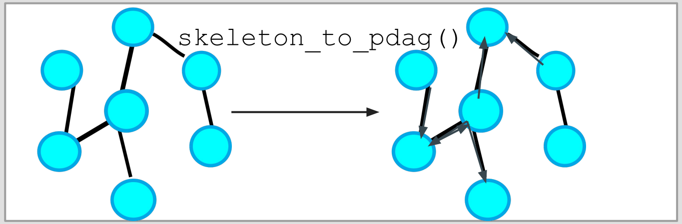
(Fukumizu, 2015)

The algorithm has 3 main phases:

1. **Skeleton phase**
   `build_skeleton()`
2. **Collider phase**
   `skeleton_to_pdag()`
3. **Generalized Transitive phase**
   `regrVonPS()`
   +
   `PDAG_by_Meek_rules()`

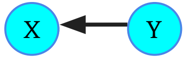
build_skeleton(data, ci_test, alpha, variant)



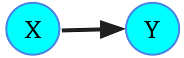


`regrVonPS()`

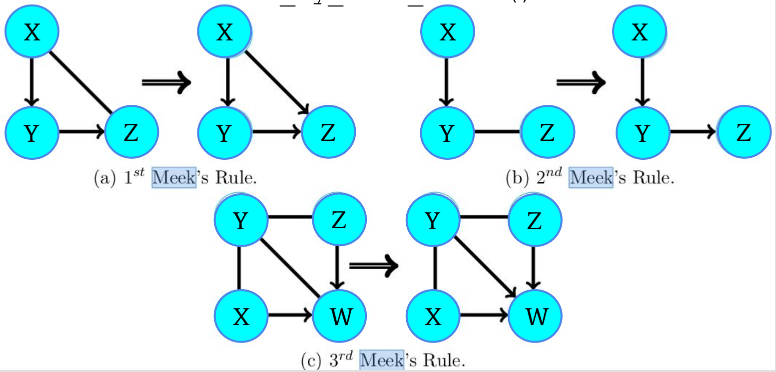
$$\begin{cases} r_X \not\perp\!\!\!\perp Y \\ r_Y \perp\!\!\!\perp X \end{cases}$$



$$\begin{cases} r_X \perp\!\!\!\perp Y \\ r_Y \not\perp\!\!\!\perp X \end{cases}$$



`PDAG_by_Meek_rules()`

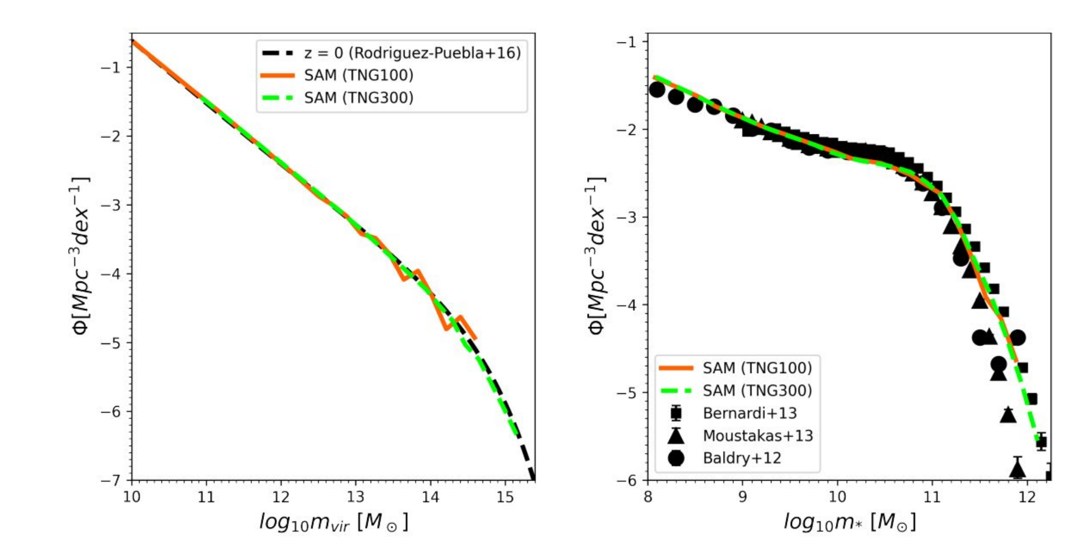


First application: what causes disk size evolution ?

simulated data from

SAM : Santa-Cruz TNG

(Gabrielpillai, 2021)



Collaborators

- V. Acquaviva (Prof @ CUNY)
- F. Bucinca (PhD graduate , CUNY)
- A. Maller (CUNY/CCA)
- R. Trotta (SISSA)

Pipeline of the analysis

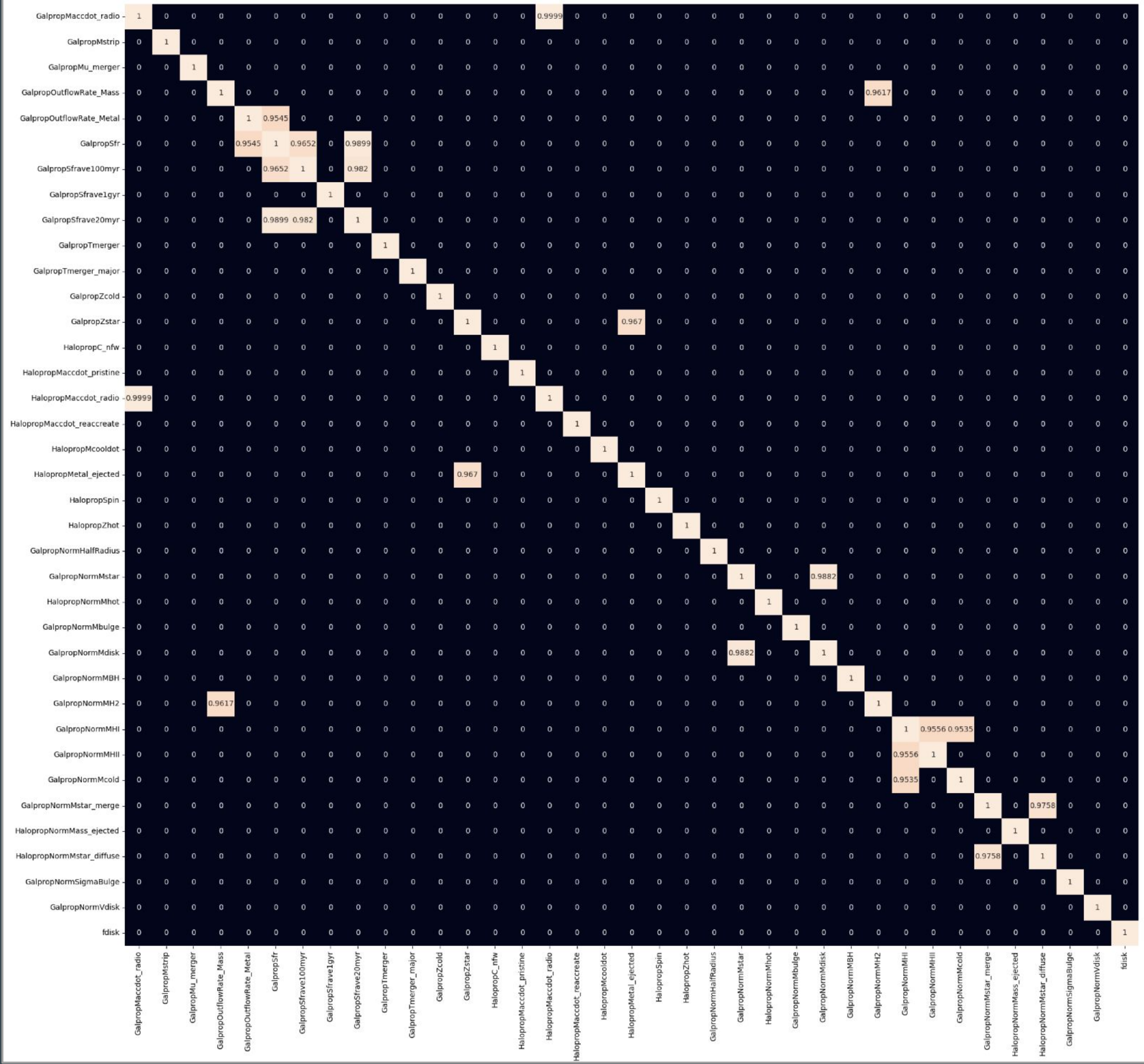
1. Simulated galaxy catalog creation
2. preprocessing
3. subsampling
4. run kernel-PC on every subsample
5. estimate average reconstructed graph with 10 bootstrap
6. apply pySR to derive quantitative laws/relations

Preprocessing

1. subsetting data by galaxy morphology (disks / ellipticals)

$$M_{bulge}/M_{star} < 0.4$$

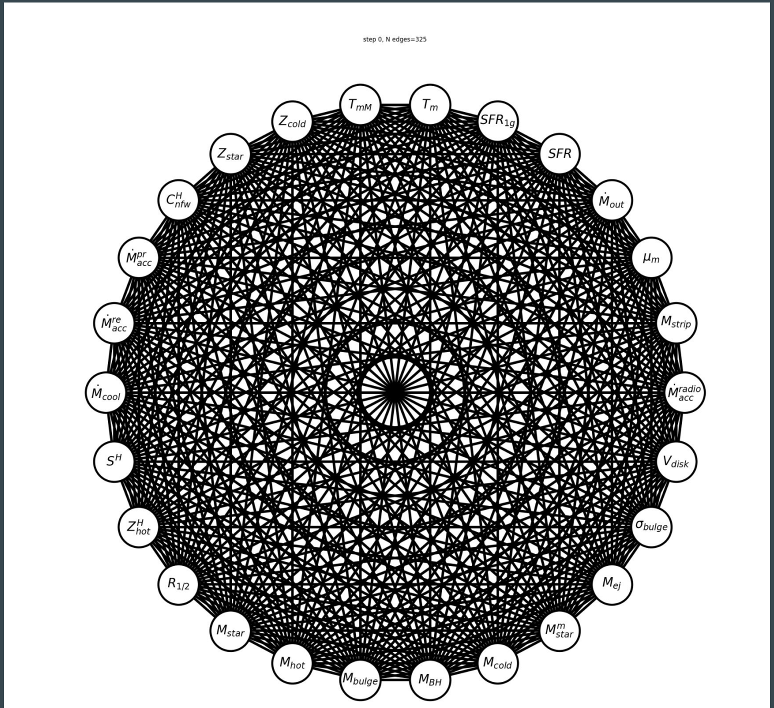
2. identification of collinear variables

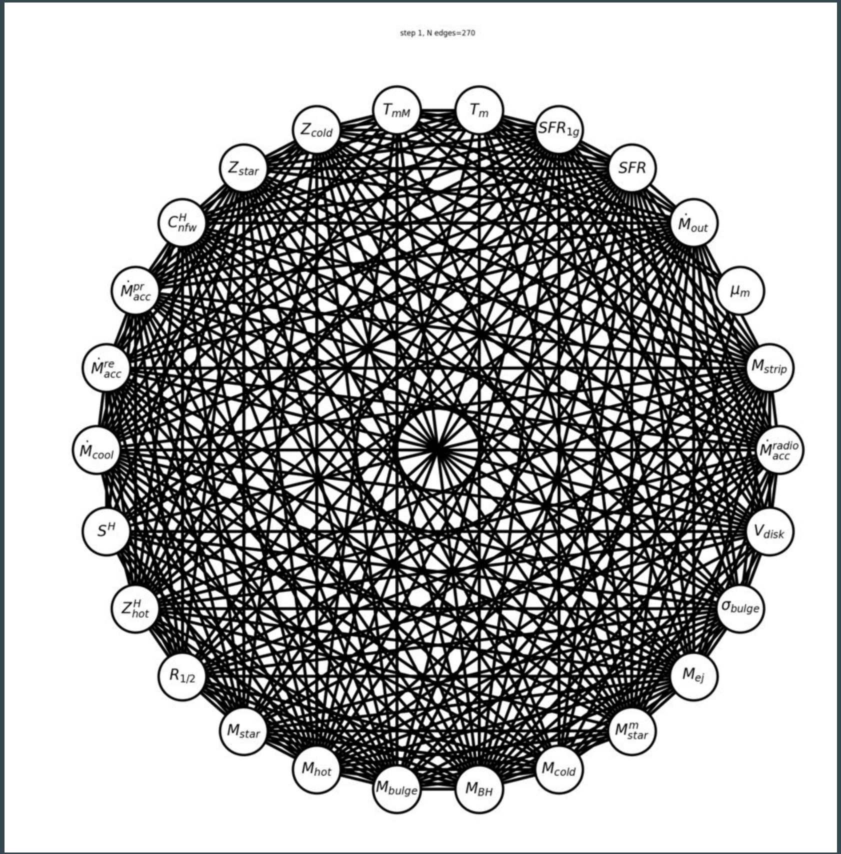


After preprocessing -> we are left with 26 vars out of 37

Name	description	Name	description
M_{strip}	stripping mass	SFR_{1g}	$< SFR >_{1Gyr}$
$\dot{M}_{outflow}$	outflow rate	T_{merge}	T last merger
μ_{merger}	mass ratio mergers	T_{mergeM}	T last major merger
Z_{cold}	metallicity cold gas	$\dot{M}^{pr}_{acc}$	pristine accretion rate
$\dot{M}^{re}_{acc}$	reaccretion rate	C_{nfw}	halo concentration
$Spin^{H}$	halo spin	$\dot{M}_{cool}$	cooling mass
$R_{1/2}$	half-mass radius	Z^{H}_{hot}	metallicity hot gas
M_{BH}	BH mass	M_{cold}	CG mass
$M_{ejected}$	Ejected Mass	V_{disk}	disk vel.
M_{star}	stellar mass	M_{hot}	Hot gas mass
σ_{bulge}	bulge vel. disp.	SFR	Star formation rate
M_{bulge}	bulge mass	Z_{star}	stellar metallicity
M^{merge}_{star}	stellar mass from merg.	$\dot{M}^{radio}_{acc}$	radio accretion rate

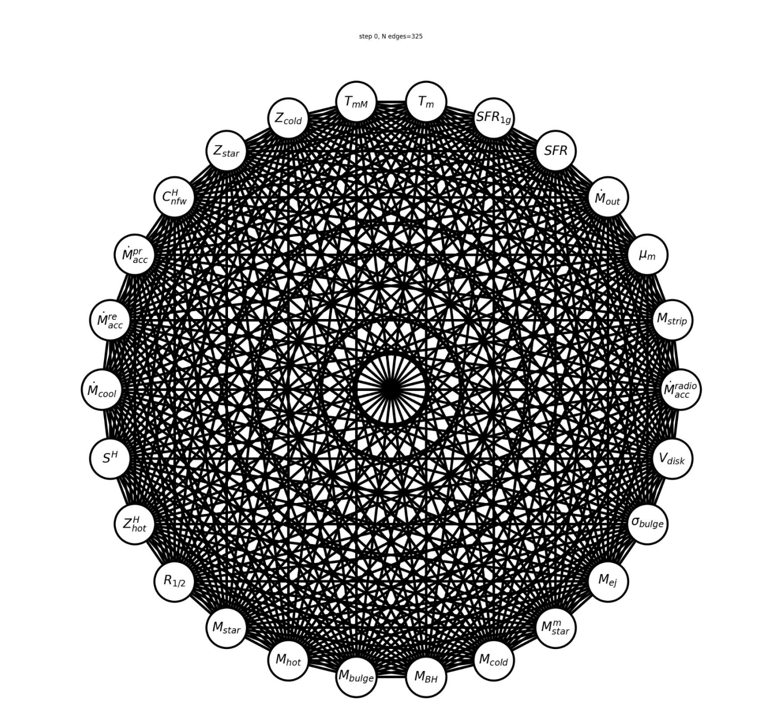
Skeleton building phase

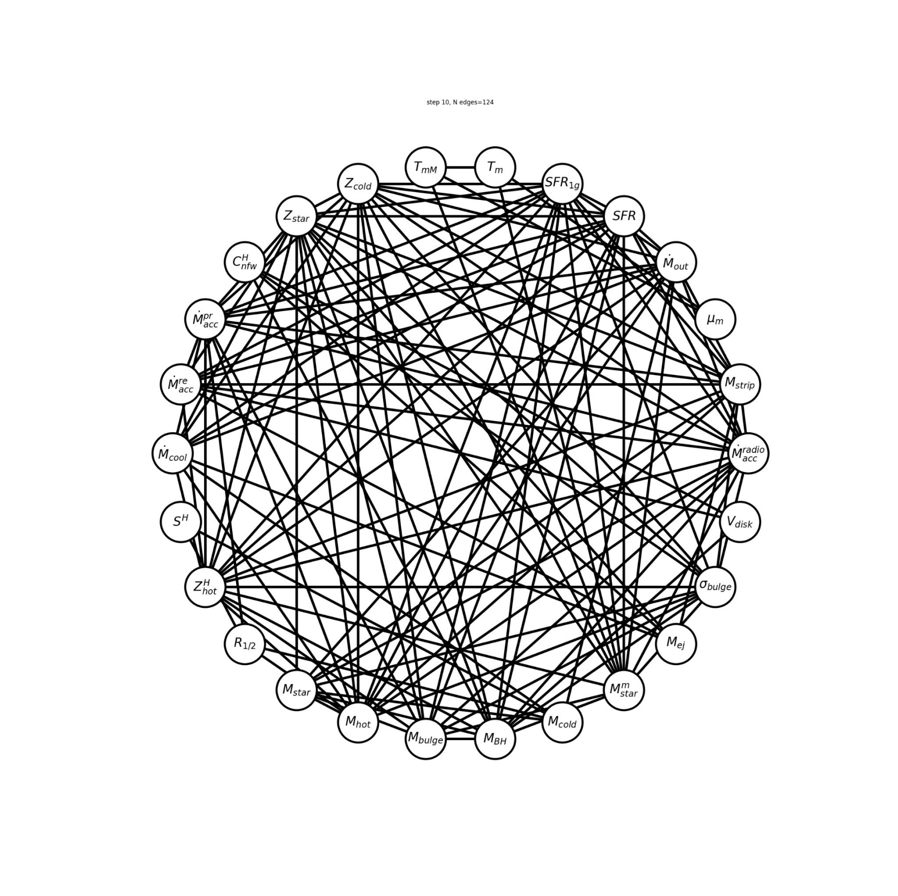




We start from the moral graph and we stop at S_set size=10

Skeleton building phase

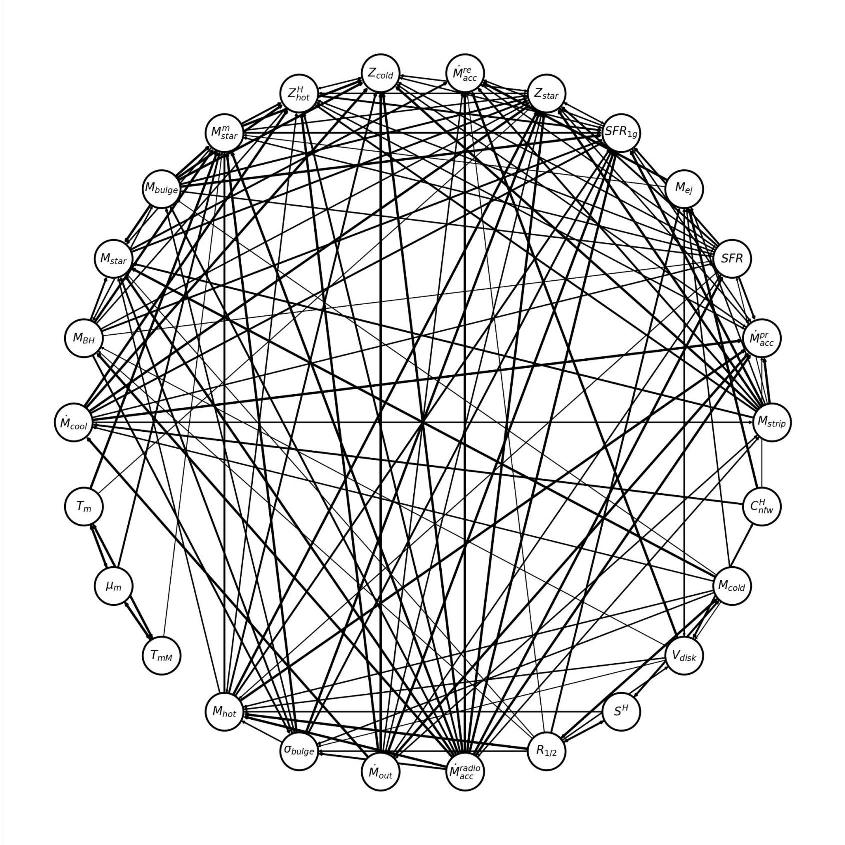




We start from the moral graph and we stop at S_set size=10

Average reconstructed BN

We adopt the approach proposed by
<u>Friedman, Goldszmidt and Wyner (1999):</u>
generating multiple network structures by applying nonparametric bootstrap to the data and estimating the relative frequency of the feature of interest



BN averaged over 10 bootstrap subsamples (N=4000)

Future research directions

After concluding the current study on causal graphical tools applied to galaxy simulated data we plan to move to real data, and we plan to integrate in the gABi package:

- treatment for latent variables (based on SEM packages, e.g. SEMOPY)
- treatment for time-series data (already under implementation)

For more general application we will add to gABi package:

- treatment for mixed categorical/continuous variables
- new CI tests: Shrinkage test
- new possible moves for the CI sampling, like the so-called new edge reversal move (Grzegorczyk
  and Husmeier, 2008)
- new structure-learning algorithms: layering-MCMC*, hill-climbing, greedy search

This will allow to perform Bayesian inference on time-series data, even in the case where latent confounders are present.

LONG-TERM GOAL: causal model discovery on real data

References

Edwards DI (2000). Introduction to Graphical Modelling. Springer, 2nd edition

Fukumizu(2007) Kernel Measures of Conditional Dependence

Tsamardinos I, Borboudakis G (2010). "Permutation Testing Improves Bayesian Network Learning". Machine Learning and Knowledge Discovery in Databases, 322–337

Tillman, R. E., Gretton, A. and Spirtes, P. (2009). Nonlinear directed acyclic structure learning with weakly additive noise model. NIPS 22, Vancouver

Colombo D, Maathuis MH (2014). "Order-Independent Constraint-Based Causal Structure Learning". Journal of Machine Learning Research, 15:3921–3962. (article introducing pc-stable)

Matthew J. Wovels et al. (2021), D'ya like DAGs? A Survey on Structure Learning and Causal Discovery, arXiV 2103.02582v2

Russell SJ, Norvig P (2009). Artificial Intelligence: A Modern Approach. Prentice Hall, 3rd edition

The End