Causal inference from interventional data

Alain Hauser

Department of Biology, Bioinformatics, University of Bern

December 10, 2013, Angers

Joint work with Peter Bühlmann

Example of a causal question

- People with sleep problems tend to be more depressed than people without sleep problems
- Do sleep problems cause depression?

Example of a causal question

- People with sleep problems tend to be more depressed than people without sleep problems
- Do sleep problems cause depression?

Possible scenarios:

sleep problems $\longrightarrow$ depression

sleep problems $\longleftarrow$ depression

$$\begin{tikzcd}[column sep=small] & \text{stress} \arrow[dl] \arrow[dr] & \\ \text{sleep problems} & & \text{depression} \end{tikzcd}$$

Example of a causal question

- People with sleep problems tend to be more depressed than people without sleep problems
- Do sleep problems cause depression?

Possible scenarios:

$$\text{sleep problems} \longrightarrow \text{depression}$$

$$\text{sleep problems} \longleftarrow \text{depression}$$

$$\begin{tikzcd} & \text{stress} \arrow[dl] \arrow[dr] & \\ \text{sleep problems} & & \text{depression} \end{tikzcd}$$

Assumptions:

- No hidden variables

Example of a causal question

- People with sleep problems tend to be more depressed than people without sleep problems
- Do sleep problems cause depression?

Possible scenarios:

$$\boxed{\begin{array}{l} \text{sleep problems} \longrightarrow \text{depression} \\ \text{sleep problems} \longleftarrow \text{depression} \end{array}}$$

$$\cancel{\text{sleep problems} \longleftarrow \text{stress} \longrightarrow \text{depression}}$$

Assumptions:

- No hidden variables
- No cyclic dependencies

Causal inference: motivated by biology

- Does depression cause sleep problems?
- Does a certain drug cure sleep problems?

Causal inference: motivated by biology

- Does depression cause sleep problems?
- Does a certain drug cure sleep problems?
- Which proteins regulate the expression of a specific gene?

Causal inference: motivated by biology

- Does depression cause sleep problems?
- Does a certain drug cure sleep problems?
- Which proteins regulate the expression of a specific gene?
- Type of regulation: inhibition, activation?
- Strength of effect?

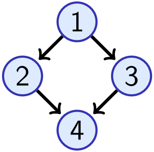
Causal inference: motivated by biology

- Does depression cause sleep problems?
- Does a certain drug cure sleep problems?
- Which proteins regulate the expression of a specific gene?
- Type of regulation: inhibition, activation?
- Strength of effect?

Aim: detection of **causal networks** modelled by **directed acyclic graphs** (DAGs)

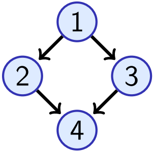
Causal model: example

Directed acyclic graph (**DAG**) D of causal dependencies:



Causal model: example

Directed acyclic graph (**DAG**) D of causal dependencies:



Random variables $X_1, \ldots, X_4$: expression levels of 4 genes

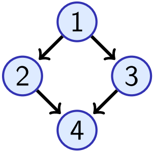
Joint density

$$f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$$

f has **Markov property** of D

Causal model: example

Directed acyclic graph (**DAG**) D of causal dependencies:



Random variables $X_1, \ldots, X_4$: expression levels of 4 genes

Joint density

$$f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$$

f has **Markov property** of D

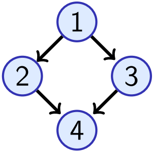
Statements encoded in causal model

- Conditional independence relations between random variables (**Markov property**)
- Effects of forcing random variables to chosen values (**intervention effects**)

Intervention: example

Random variables:

X_1: exp. level of gene 1
X_2: exp. level of gene 2
X_3: exp. level of gene 3
X_4: exp. level of gene 4



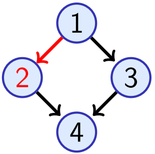
True DAG D

Observational density: $f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$

Intervention: example

Random variables:

X_1: exp. level of gene 1
X_2: exp. level of gene 2
X_3: exp. level of gene 3
X_4: exp. level of gene 4



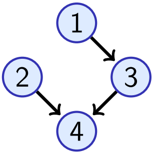
Intervention at X_2: silencing gene 2

Observational density: $f(x) = f(x_1) f(x_2|x_1) f(x_3|x_1) f(x_4|x_2, x_3)$

Intervention: example

Random variables:

X_1: exp. level of gene 1
X_2: exp. level of gene 2
X_3: exp. level of gene 3
X_4: exp. level of gene 4



Intervention DAG $D^{(\{2\})}$

Observational density: $f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$

Interventional density: $f(x|\mathrm{do}(X_2 = U)) = f(x_1)\tilde{f}(x_2)f(x_3|x_1)f(x_4|x_2, x_3)$

Example: estimating effect of gene knockouts in yeast

(Maathuis et al., 2010)

- $n = 63$ measurements of $X_1, \ldots, X_p$ ($p = 5361$): gene expression levels in yeast
- **Question:** which genes are strongly affected by the knockout of other genes?

Example: estimating effect of gene knockouts in yeast

- **"Classical" approach:** regression: $X_i = \sum_{j \neq i} \beta_j X_j + \varepsilon$

  $|\beta_j|$ measures change of X_i as function of X_j when **keeping all other variables fixed**.

Example: estimating effect of gene knockouts in yeast

- **"Classical" approach:** regression: $X_i = \sum_{j \neq i} \beta_j X_j + \varepsilon$

  $|\beta_j|$ measures change of X_i as function of X_j when **keeping all other variables fixed**.
- Not very realistic
  - complex interplay between genes of an organism
  - silencing one gene affects many others
  - indirect regulation paths should be accounted for

Example: estimating effect of gene knockouts in yeast

- **"Classical" approach:** regression: $X_i = \sum_{j \neq i} \beta_j X_j + \varepsilon$

  $|\beta_j|$ measures change of X_i as function of X_j when **keeping all other variables fixed**.
- Not very realistic
  - complex interplay between genes of an organism
  - silencing one gene affects many others
  - indirect regulation paths should be accounted for
- **Causal approach:**
  - estimate directed acyclic graph (DAG) of direct influences
  - graph as a whole can also model **indirect** influences
  - more realistic scenario

Example: estimating effect of gene knockouts in yeast

Data set of Hughes et al. (2000): expression levels of 5361 yeast genes, originating from...

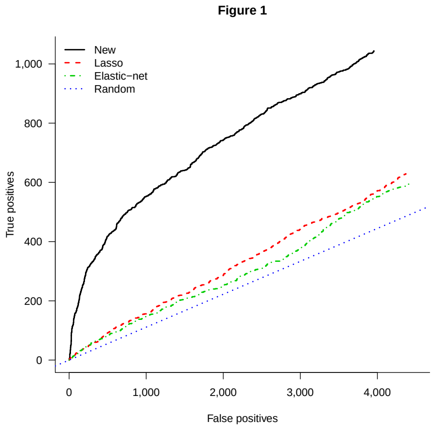
- 63 wildtype cells
- 234 mutants

Procedure of Maathuis et al. (2010):

- "Knockout effect": difference in expression of one gene in response to knockout of another gene
- Find strongest 5% of "knockout effects" in mutants data
- Predict strongest α% of knockout effects based on model fitted to wildtype data
- Compare predictions of different methods with ROC curves

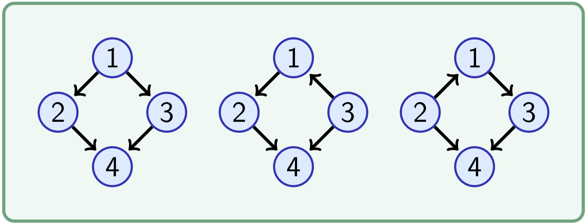
Example: estimating effect of gene knockouts in yeast

Indeed: causal method outperforms classical regression models!



Markov equivalence

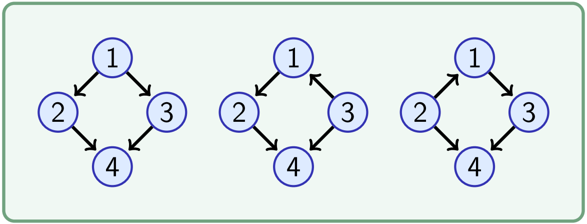
A probability density in general obeys the Markov properties of **several** DAGs; those DAGs are called **Markov equivalent**
$\rightsquigarrow$ **limited identifiability** under observational data



D D_1 D_2

Markov equivalence

A probability density in general obeys the Markov properties of **several** DAGs; those DAGs are called **Markov equivalent**
$\rightsquigarrow$ **limited identifiability** under observational data



D D_1 D_2

On the other hand, intervention effects **do** depend on the DAG
$\rightsquigarrow$ **improved identifiability** of causal models under interventional data

Interventional Markov equivalence

Definition (Interventional Markov equivalence)

Two DAGs D_1 and D_2 are **interventionally Markov equivalent** *for a given set of intervention targets* if they

- encode the same interventional densities
- are statistically indistinguishable under intervention experiments performed at the specified intervention targets.

Interventional Markov equivalence

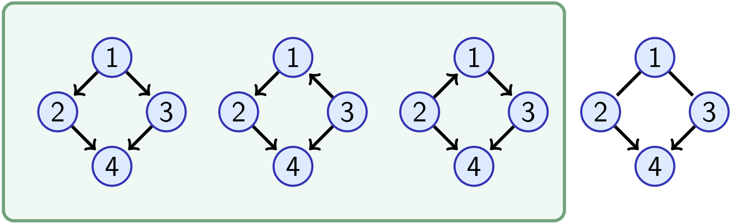
Definition (Interventional Markov equivalence)

Two DAGs D_1 and D_2 are **interventionally Markov equivalent** *for a given set of intervention targets* if they

- encode the same interventional densities
- are statistically indistinguishable under intervention experiments performed at the specified intervention targets.

- Observational setting is a special case of an interventional setting
- $\exists$ purely graph theoretic criterion for interventional Markov equivalence (Hauser and Bühlmann, 2012)
- Reproduces classical criterion for observational Markov equivalence of Verma and Pearl (1990):
  DAGs D_1 and D_2 observationally Markov equivalent $\Leftrightarrow$ D_1 and D_2 have same skeleton and v-structures.

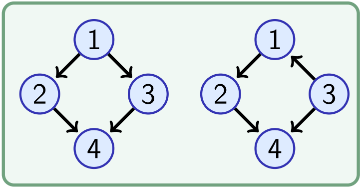
Interventional Markov equivalence: example



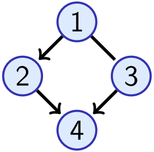
D $\quad D_1 \quad D_2 \quad \mathcal{E}_{\{\emptyset\}}(D)$

Observational Markov equivalence class of D with corresponding essential graph

Interventional Markov equivalence: example



D D_1



$\mathcal{E}_{\{\emptyset,\{2\}\}}(D)$

Interventional Markov equivalence class of D assuming we can measure

- observational data
- interventional data from an intervention at X_2

Interventional essential graph

Interventional essential graph $\mathcal{E}_{\mathcal{I}}(D)$ of a DAG D: partially directed graph

- having the same skeleton as D
- with a **directed edge** where the corresponding arrows of all DAGs interventionally equivalent to D have the same orientation
- with an **undirected edge** where the orientation of the corresponding arrow is *not* common to all DAGs interventionally equivalent to D

$\mathcal{I}$: set of intervention targets

Interventional essential graph: unique representation of interventional Markov equivalence class

Characterization of $\mathcal{I}$-essential graphs

Theorem (Hauser and Bühlmann, 2012)

A graph G is the $\mathcal{I}$-essential graph of a DAG D if and only if

1. *G is a chain graph;*
2. *each chain component of G is chordal;*
3. *$a \longrightarrow b \longrightarrow c$ is no induced subgraph of G;*
4. *G has no line $a \longrightarrow b$ for which there exists some $I \in \mathcal{I}$ such that $|I \cap \{a, b\}| = 1$;*
5. *every arrow $a \longrightarrow b \in G$ is strongly $\mathcal{I}$-protected.*

Reproduces a result of Andersson et al. (1997) for the observational case $\mathcal{I} = \{\emptyset\}$.

Characterization of $\mathcal{I}$-essential graphs

Theorem (Hauser and Bühlmann, 2012)

A graph G is the $\mathcal{I}$-essential graph of a DAG D if and only if

1. *G is a chain graph;*
2. *each chain component of G is ch*
3. *$a \longrightarrow b — c$ is no induced ... of G;*
4. *G has no line $a — b$ ... ich there exists some $I \in \mathcal{I}$ such that $|I \cap \{a, b\}| = 1$;*
5. *every arrow $a \longrightarrow b \in G$ is strongly $\mathcal{I}$-protected.*

technical detail

Reproduces a result of Andersson et al. (1997) for the observational case $\mathcal{I} = \{\emptyset\}$.

Interventional Markov equivalence: summary

- Causal models not fully identifiable from observational data
- Interventional data improves identifiability

Interventional Markov equivalence: summary

- Causal models not fully identifiable from observational data
- Interventional data improves identifiability
- Graph theoretic criterion for interventional Markov equivalence of two DAGs
- Interventional essential graphs: representation of $\mathcal{I}$-Markov equivalence classes for visualization and algorithmic handling

Interventional Markov equivalence: summary

- Causal models not fully identifiable from observational data
- Interventional data improves identifiability
- Graph theoretic criterion for interventional Markov equivalence of two DAGs
- Interventional essential graphs: representation of $\mathcal{I}$-Markov equivalence classes for visualization and algorithmic handling

- Next part: learning $\mathcal{I}$-equivalence classes from data

Gaussian causal model

- Gaussian causal model: $X \sim \mathcal{N}(0, \Sigma)$; density has Markov property of some DAG D
- Markov property translates to a set of **linear** structural equations:

$$X_k = \sum_{k=1}^{p} \beta_{kj} X_j + \varepsilon_k, \quad \varepsilon_k \overset{\text{indep.}}{\sim} \mathcal{N}(0, \sigma_k^2), \quad 1 \leq k \leq p$$

with $\beta_{kj} = 0$ if there is no arrow from j to k in the DAG D.

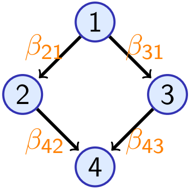
Gaussian causal model

- Gaussian causal model: $X \sim \mathcal{N}(0, \Sigma)$; density has Markov property of some DAG D
- Markov property translates to a set of **linear** structural equations:

$$X_k = \sum_{k=1}^{p} \beta_{kj} X_j + \varepsilon_k, \quad \varepsilon_k \overset{\text{indep.}}{\sim} \mathcal{N}(0, \sigma_k^2), \quad 1 \le k \le p$$

with $\beta_{kj} = 0$ if there is no arrow from j to k in the DAG D.

- Family of models parameterized by the "edge weights" $B := (\beta_{kj})_{k,j=1}^{p}$ and the error variances $\sigma^2 := (\sigma_1^2, \ldots, \sigma_p^2)$.



Likelihood for given DAG

- Calculation of **maximum likelihood estimator** (MLE) for edge weights $\hat{B}$ and error variances $\hat{\sigma^2}$ for **jointly observational and interventional data**: decouples into optimization over single structural equations
- $(\hat{\beta}_{kj})_{j=1}^{p}$, $\hat{\sigma}_k^2$: given by least-squares regression of $X_k \sim X_{\mathrm{pa}(k)}$ (measurements of one variable vs. its "parents"), ignoring samples produced by intervention at X_k (Hauser and Bühlmann, 2013)

Likelihood for given DAG

- Calculation of **maximum likelihood estimator** (MLE) for edge weights $\hat{B}$ and error variances $\hat{\sigma^2}$ for **jointly observational and interventional data**: decouples into optimization over single structural equations
- $(\hat{\beta}_{kj})_{j=1}^{p}$, $\hat{\sigma}_k^2$: given by least-squares regression of $X_k \sim X_{\mathrm{pa}(k)}$ (measurements of one variable vs. its "parents"), ignoring samples produced by intervention at X_k (Hauser and Bühlmann, 2013)
  $\rightsquigarrow$ **parameter estimation:** analytical calculation of MLE
  $\rightsquigarrow$ **model selection:** efficient calculation of **Bayesian information criterion** (BIC)

Learning causal models

- For fix p, optimization of the BIC leads to **consistent** model selection in the limit $n \to \infty$ (Hauser and Bühlmann, 2013)

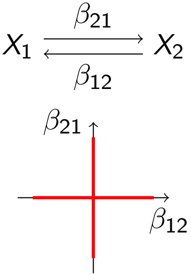
Learning causal models

- For fix p, optimization of the BIC leads to **consistent** model selection in the limit $n \to \infty$ (Hauser and Bühlmann, 2013)
- Problem: **model selection** by optimizing BIC is computationally intrinsically hard (NP-hard; Chickering, 1996)

Learning causal models

- For fix p, optimization of the BIC leads to **consistent** model selection in the limit $n \to \infty$ (Hauser and Bühlmann, 2013)
- Problem: **model selection** by optimizing BIC is computationally intrinsically hard (NP-hard; Chickering, 1996)

- Replacing ℓ_0 by ℓ_1 regularization does not help; reason: **DAG constraint** (non-convex constraint!)

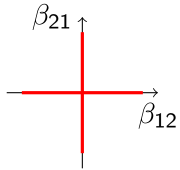


Learning causal models

- For fix p, optimization of the BIC leads to **consistent** model selection in the limit $n \to \infty$ (Hauser and Bühlmann, 2013)
- Problem: **model selection** by optimizing BIC is computationally intrinsically hard (NP-hard; Chickering, 1996)

- Replacing ℓ_0 by ℓ_1 regularization does not help; reason: **DAG constraint** (non-convex constraint!)
- **Solution:** causal inference via **greedy algorithm** on space of $\mathcal{I}$-essential graphs $\rightsquigarrow$ Greedy Interventional Equivalence Search (GIES): natural generalization of the Greedy Equivalence Search (GES) algorithm of Chickering (2002) to interventional data

$$X_1 \underset{\beta_{12}}{\overset{\beta_{21}}{\rightleftarrows}} X_2$$



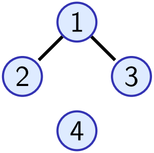
GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

Possible forward step:

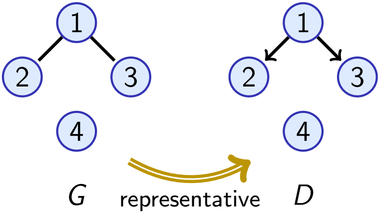


G

GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

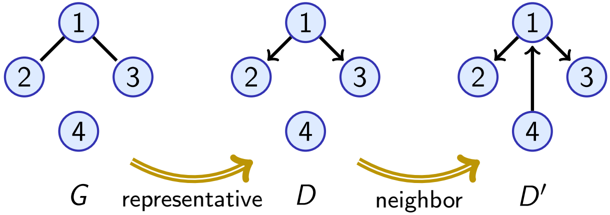
Possible forward step:



GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

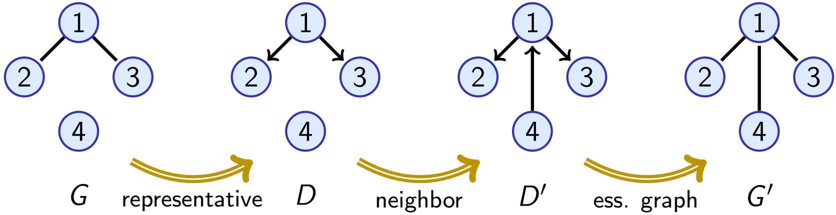
Possible forward step:



GIES: example step

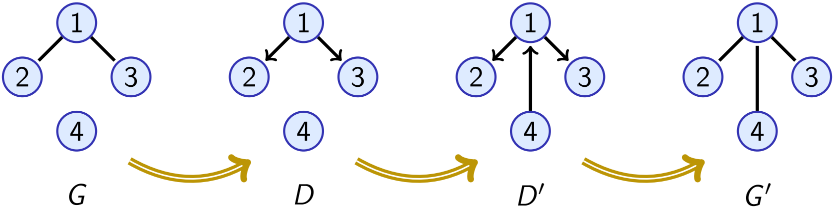
- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

Possible forward step:



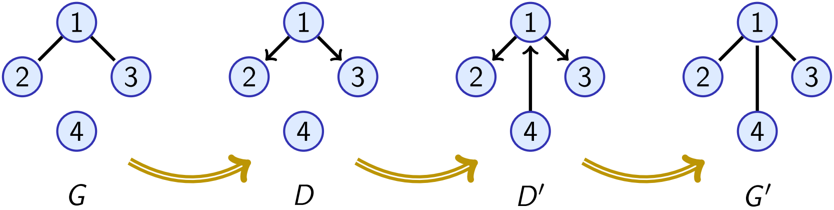
Search space: DAGs vs. essential graphs

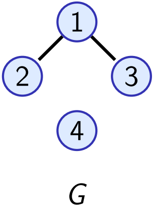
Neglecting (interventional) Markov equivalence narrows search space



Search space: DAGs vs. essential graphs

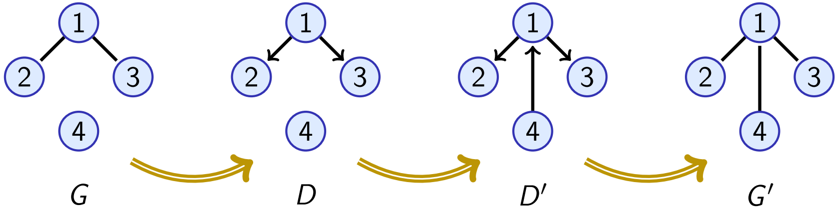
Neglecting (interventional) Markov equivalence narrows search space

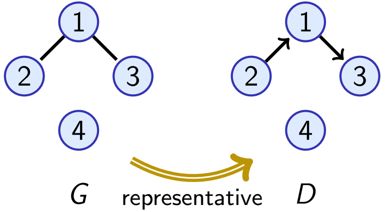




Search space: DAGs vs. essential graphs

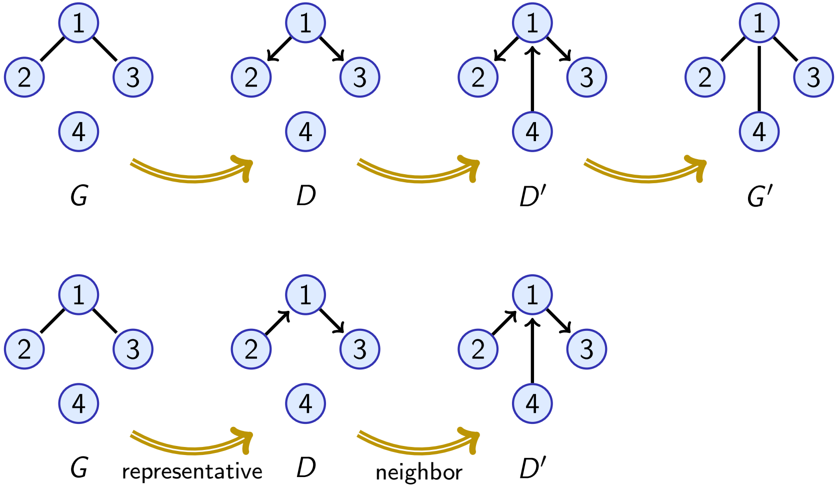
Neglecting (interventional) Markov equivalence narrows search space





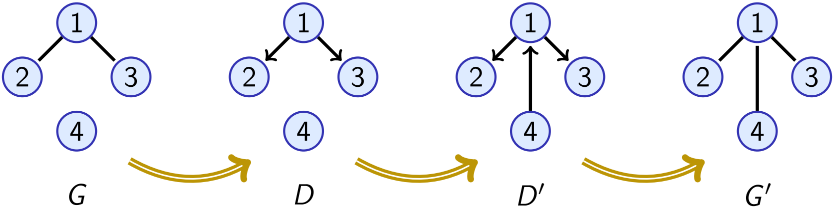
Search space: DAGs vs. essential graphs

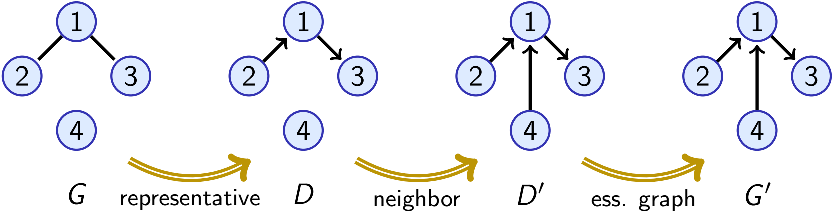
Neglecting (interventional) Markov equivalence narrows search space



Search space: DAGs vs. essential graphs

Neglecting (interventional) Markov equivalence narrows search space



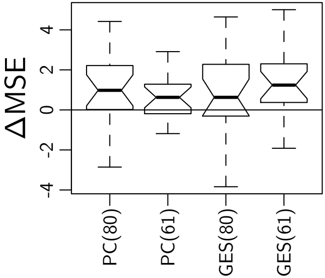


DREAM4 *in silico* network challenge

- **Goal**: learn structure of gene regulatory network, predict intervention effects
- **Data**: realistic *in silico* steady-state and time series data, observational and interventional data points
- Our proceeding: **cross-validation** of gene expression levels under interventions.
- Compare CV-values to those of algorithms ignoring interventional nature of data

DREAM4 challenge: results

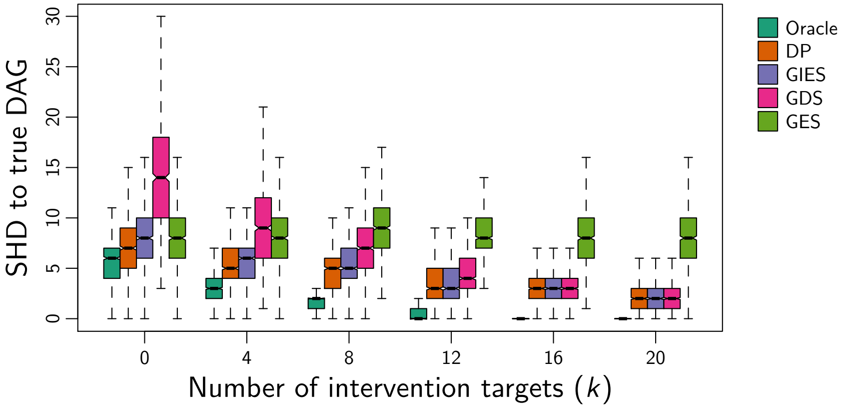
ΔMSE := MSE of competitor − MSE of GIES



Conclusions:

- slight advantage over competing methods
- estimation sensitive to model misspecification: acyclicity and normality assumptions violated

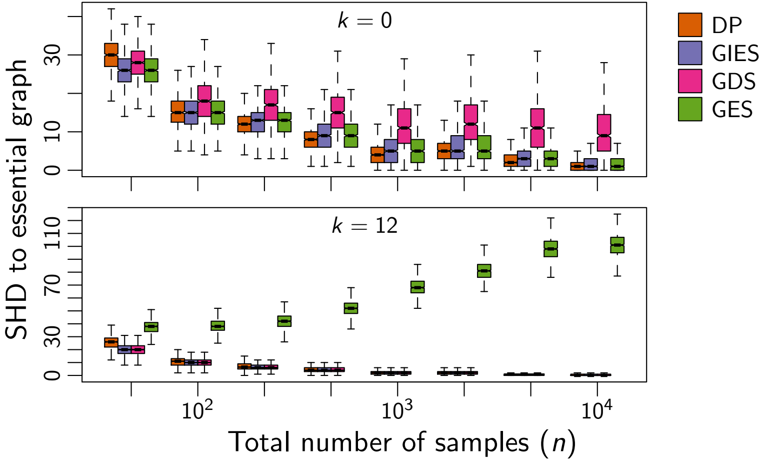
Simulation study: structure learning



Structural Hamming distance between true DAG and estimated interventional essential graph ($n = 1000$, $p = 20$).

Structural Hamming distance (SHD): number of edges to be added, removed, or reversed to get from one graph to a different one.

Simulation study: structure learning



SHD between estimated and true interventional essential graphs ($p = 20$). Upper part: observational data; lower part: $k = 12$ intervention targets of size 4.

Learning causal models: summary

- Gaussian causal models: analytical calculation of MLE for given DAG; p independent regression problems
- Consistent model selection (structure learning) through maximization of BIC

Learning causal models: summary

- Gaussian causal models: analytical calculation of MLE for given DAG; p independent regression problems
- Consistent model selection (structure learning) through maximization of BIC
- Structure learning computationally feasible with greedy algorithm

Learning causal models: summary

- Gaussian causal models: analytical calculation of MLE for given DAG; p independent regression problems
- Consistent model selection (structure learning) through maximization of BIC
- Structure learning computationally feasible with greedy algorithm
- Greedy algorithm keeps up with dynamic programming solution at much lower computational cost
- Neglection of interventional Markov equivalence leads to worse structure learning

Outlook and future work

- Estimators suitable for high-dimensional data

Outlook and future work

- Estimators suitable for high-dimensional data
- More complex (and hence realistic) models:
  - nonlinear dependence of a variable from its causal parents
  - cyclic models
  - time series data

Outlook and future work

- Estimators suitable for high-dimensional data
- More complex (and hence realistic) models:
  - nonlinear dependence of a variable from its causal parents
  - cyclic models
  - time series data
- Accounting for hidden variables, confounders, etc.

Outlook and future work

- Estimators suitable for high-dimensional data
- More complex (and hence realistic) models:
  - nonlinear dependence of a variable from its causal parents
  - cyclic models
  - time series data
- Accounting for hidden variables, confounders, etc.

Merci pour votre attention !

References I

S.A. Andersson, D. Madigan, and M.D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Stat.*, 25(2):505–541, 1997.

D.M. Chickering. Learning Bayesian networks is NP-complete. *Learning from data: Artificial intelligence and statistics V*, 112:121–130, 1996.

D.M. Chickering. Optimal structure identification with greedy search. *JMLR*, 3(3): 507–554, 2002.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *JMLR*, 13:2409–2464, 2012.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of corresponding Markov equivalence classes of directed acyclic graphs. *Submitted*, 2013.

T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. Functional Discovery via a Compendium of Expression Profiles. *Cell*, 102(1):109–126, 2000. ISSN 0092-8674.

M.H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

References II

T. Verma and J. Pearl. On the equivalence of causal models. In *UAI 1990*, pages 220–227, 1990.