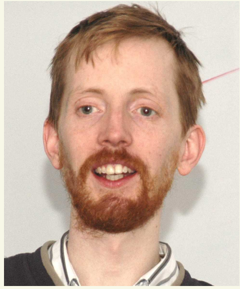
High-dimensional graphical models and causal inference (part I)

Marloes Maathuis
ETH Zurich

Collaborators

Diego Colombo



Thomas Richardson

Markus Kalisch

Peter Bühlmann

Example

- Hypothetical data on a rehabilitation program for prisoners:

	rearrested	not rearrested	rearrest rate
participants	100	400	20%
non-participants	500	500	50%

Example

- Hypothetical data on a rehabilitation program for prisoners:

	rearrested	not rearrested	rearrest rate
participants	100	400	20%
non-participants	500	500	50%

- Questions:
  1. Can we predict whether a prisoner will be rearrested based on participation in the program (and other variables)?
  2. Does the program lower the rearrest rate?
  3. What would be the rearrest rate if the program were compulsory for all prisoners?

Example

- Hypothetical data on a rehabilitation program for prisoners:

	rearrested	not rearrested	rearrest rate
participants	100	400	20%
non-participants	500	500	50%

- Questions:
  1. Can we predict whether a prisoner will be rearrested based on participation in the program (and other variables)?
  2. Does the program lower the rearrest rate? (causal mechanism)
  3. What would be the rearrest rate if the program were compulsory for all prisoners? (prediction after an intervention)

causal

Example

- Hypothetical data on a rehabilitation program for prisoners:

	rearrested	not rearrested	rearrest rate
participants	100	400	20%
non-participants	500	500	50%

- Questions:
  1. Can we predict whether a prisoner will be rearrested based on participation in the program (and other variables)?
  2. Does the program lower the rearrest rate? (causal mechanism)
  3. What would be the rearrest rate if the program were compulsory for all prisoners? (prediction after an intervention)

causal

- We need to know how the data were gathered:

Example

- Hypothetical data on a rehabilitation program for prisoners:

	rearrested	not rearrested	rearrest rate
participants	100	400	20%
non-participants	500	500	50%

- Questions:
  1. Can we predict whether a prisoner will be rearrested based on participation in the program (and other variables)?
  2. Does the program lower the rearrest rate? (causal mechanism)
  3. What would be the rearrest rate if the program were compulsory for all prisoners? (prediction after an intervention)

causal

- We need to know how the data were gathered:
  - Voluntary participation $\Rightarrow$ observational data $\Rightarrow$ question 1

Example

- Hypothetical data on a rehabilitation program for prisoners:

	rearrested	not rearrested	rearrest rate
participants	100	400	20%
non-participants	500	500	50%

- Questions:
  1. Can we predict whether a prisoner will be rearrested based on participation in the program (and other variables)?

causal

  2. Does the program lower the rearrest rate? (causal mechanism)
  3. What would be the rearrest rate if the program were compulsory for all prisoners? (prediction after an intervention)

- We need to know how the data were gathered:
  - Voluntary participation $\Rightarrow$ observational data $\Rightarrow$ question 1
  - Randomized participation $\Rightarrow$ experimental data $\Rightarrow$ questions 2/3

Causal questions

- Causal questions are about the mechanism behind the data or about predictions after some outside intervention

Causal questions

- Causal questions are about the mechanism behind the data or about predictions after some outside intervention
- Some examples:
  - Does smoking cause lung cancer?
  - What is the efficacy of a new drug?
  - What is the gene regulatory network of yeast?
  - What are major causes of global warming?
  - Did racial discrimination play a role in hiring processes?
  - What would be the effect of a new tax policy on economic growth?

Controlled experiments

- Causal questions are best answered by controlled experiments:
  - groups are equal except for the treatment condition
    $\Rightarrow$ any difference in outcome must be caused by the treatment
  - examples:
    - experiments in biology/physics/chemistry
    - clinical trials to test new drugs

Controlled experiments

- Causal questions are best answered by controlled experiments:
  - groups are equal except for the treatment condition
    $\Rightarrow$ any difference in outcome must be caused by the treatment
  - examples:
    - experiments in biology/physics/chemistry
    - clinical trials to test new drugs

- But sometimes such experiments are impossible, as they may be:
  - infeasible (global warming, life style choices)
  - unethical (smoking)
  - expensive / time consuming (gene knock-outs)

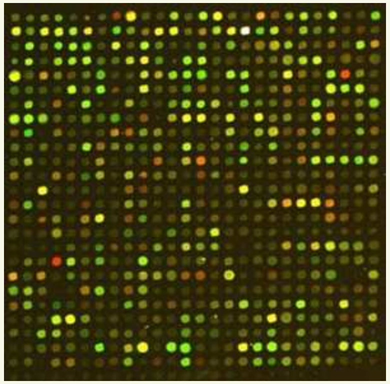
Research question

- Can we learn causal effects from observational data in high-dimensional systems?

Research question

- Can we learn causal effects from observational data in high-dimensional systems?

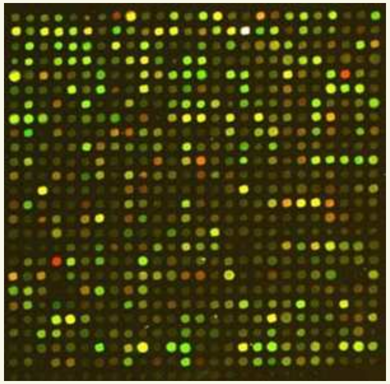
- Example: gene regulatory network of yeast:
  - identify pairs of genes between which there is a large effect from observational data
  - gene expression levels of wild-type yeast with many more variables than observations
  - > 5000 genes
  - 63 yeast organisms



Research question

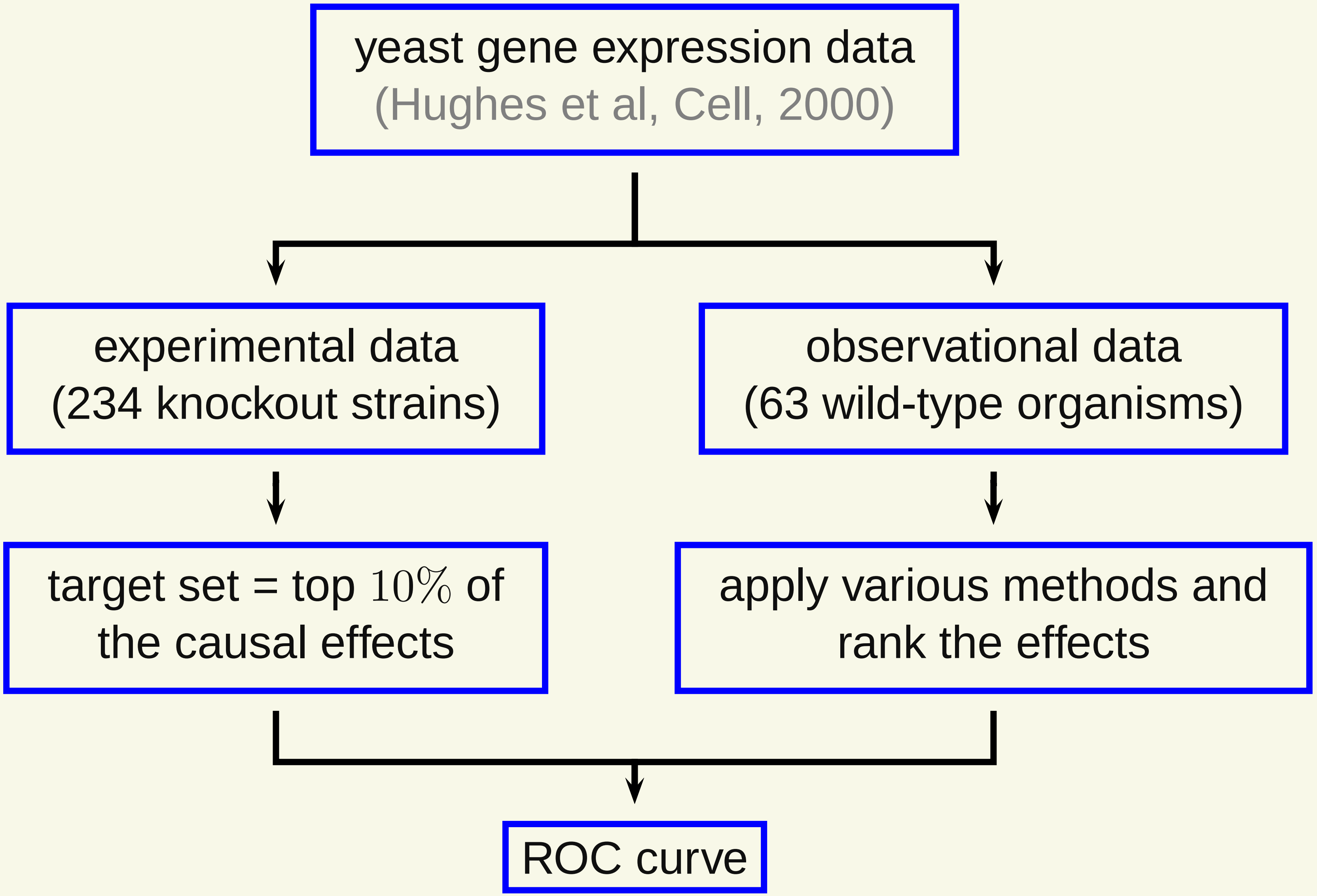
- Can we learn causal effects from observational data in high-dimensional systems?

- Example: gene regulatory network of yeast:
  - identify pairs of genes between which there is a large effect from observational data
  - gene expression levels of wild-type yeast with many more variables than observations
  - > 5000 genes
  - 63 yeast organisms

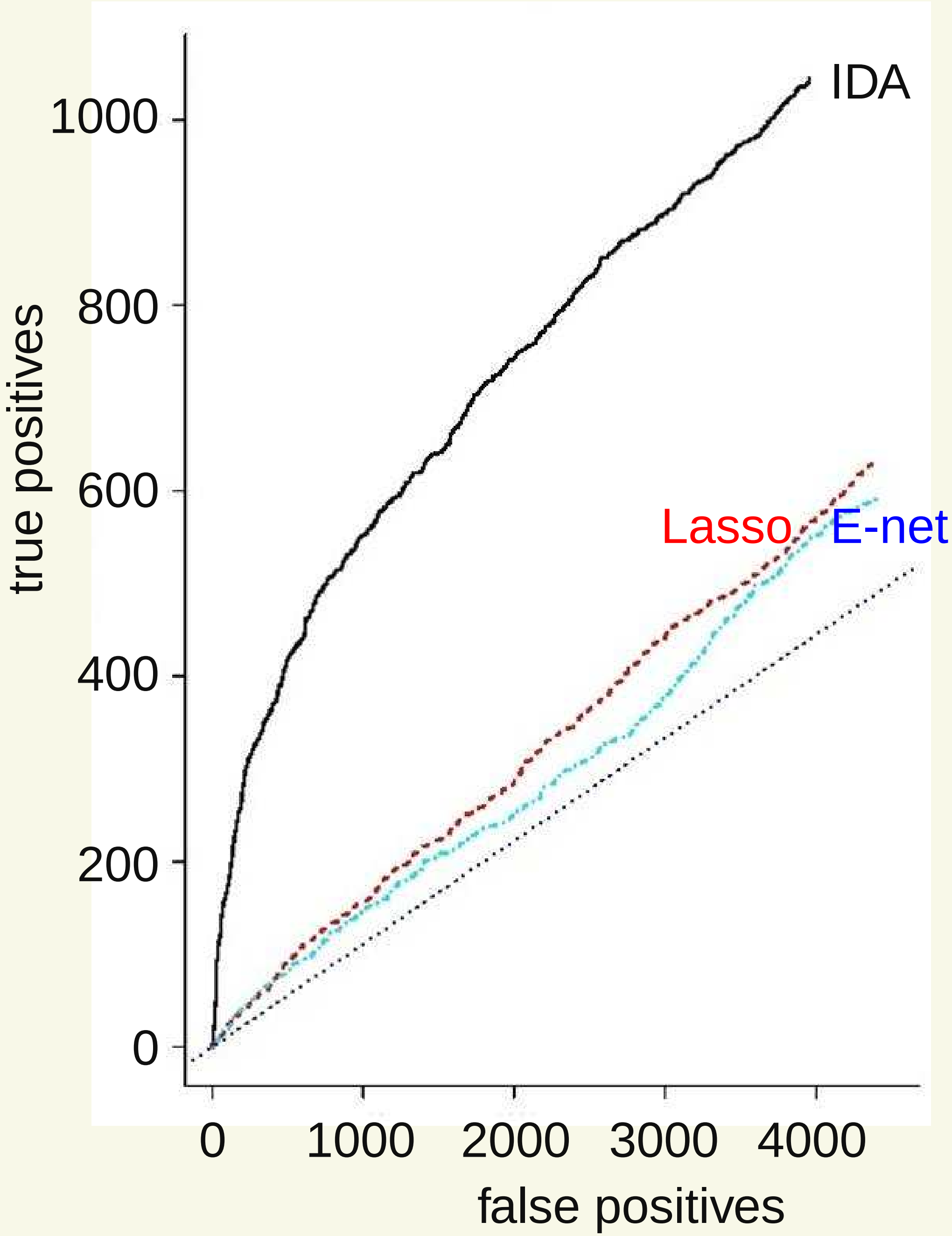


- Focus on developing scalable algorithms with known statistical properties and validations on real data

Validation

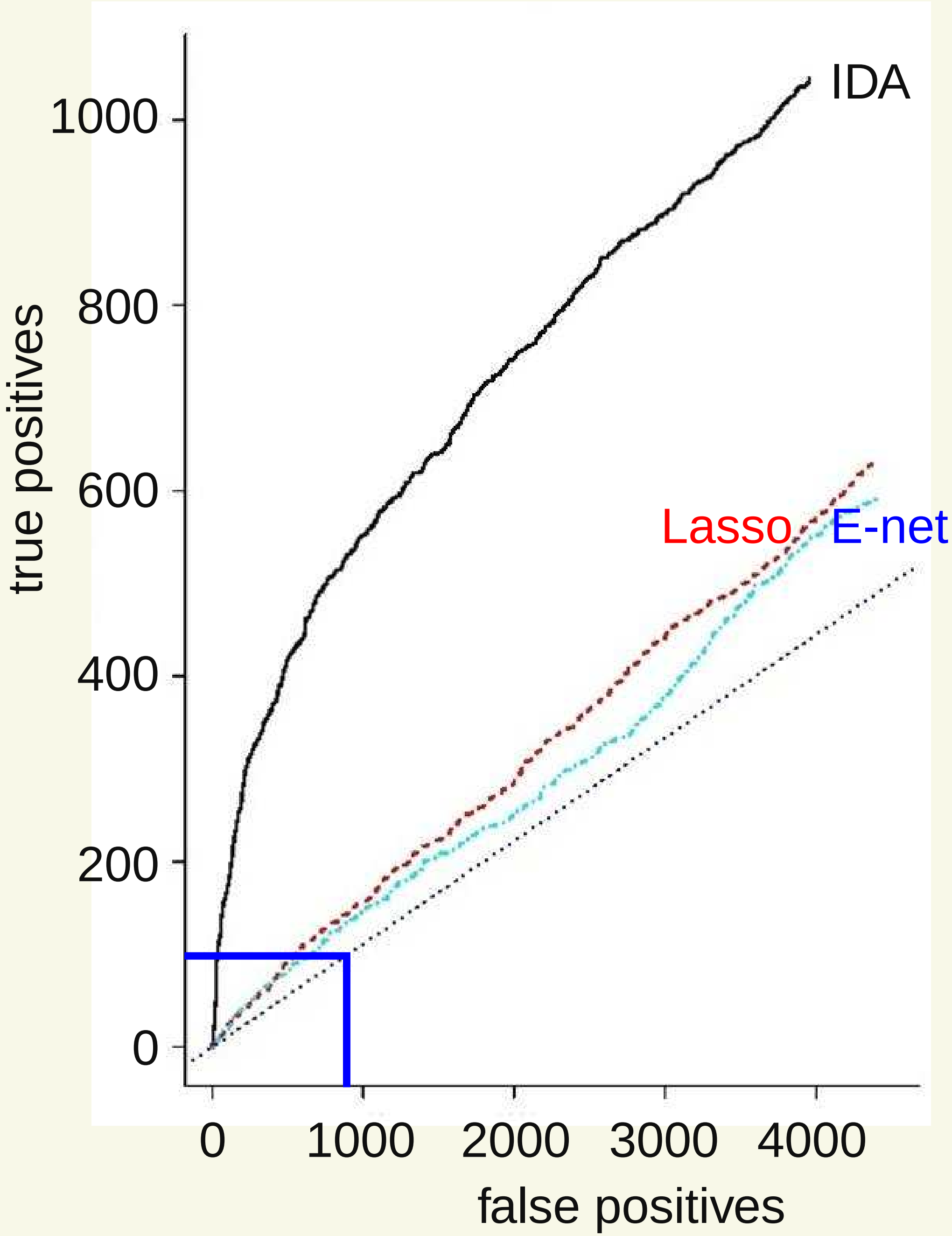

yeast gene expression data
(Hughes et al, Cell, 2000)
experimental data
(234 knockout strains)
observational data
(63 wild-type organisms)
target set = top 10% of the causal effects
apply various methods and rank the effects
ROC curve

Results (MM, Colombo, Kalisch & Bühlmann, Nature Methods, 2010)



Target set: top 10% of effects from experimental data

Results (MM, Colombo, Kalisch & Bühlmann, Nature Methods, 2010)

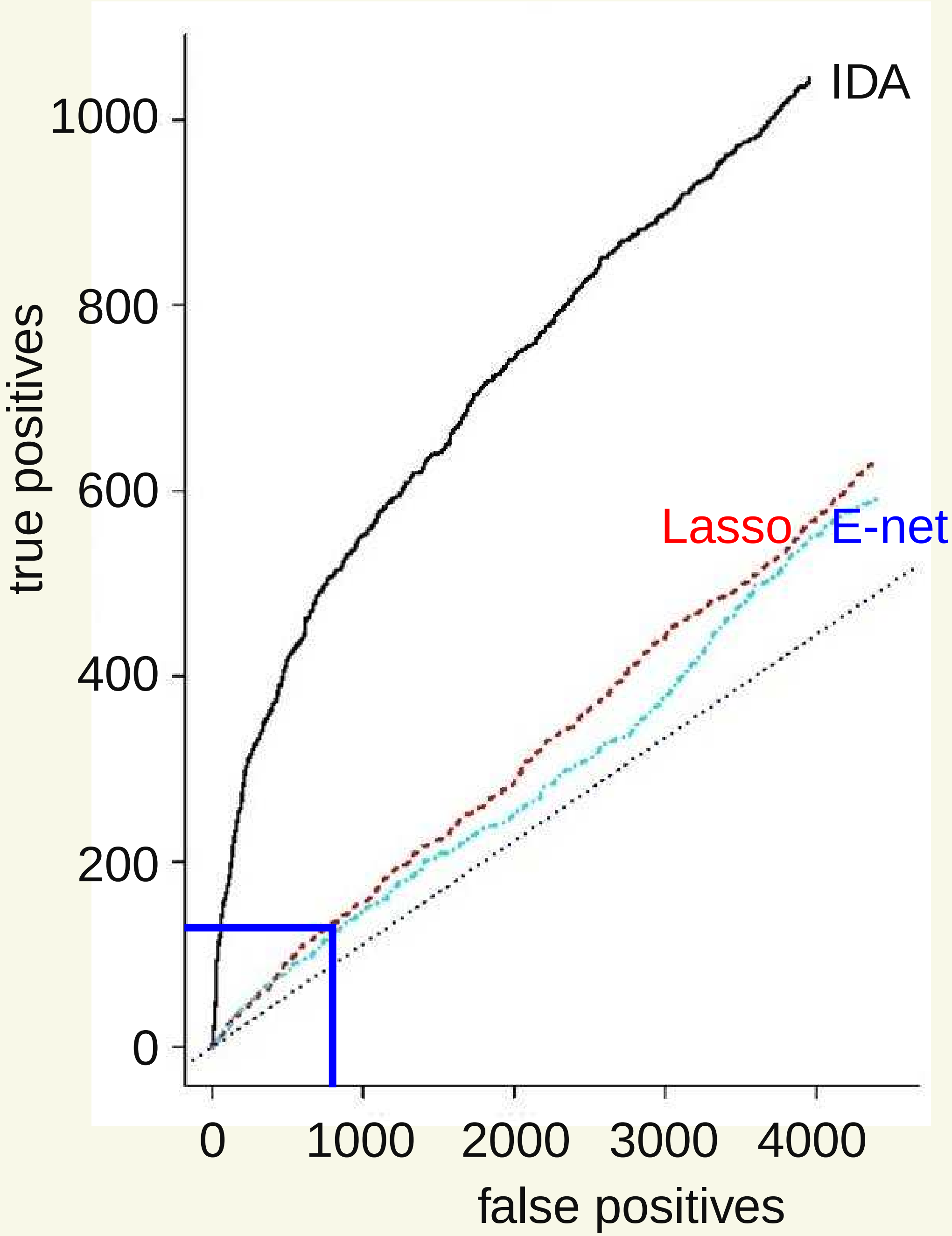


Target set: top 10% of effects from experimental data

Consider top $q = 1000$ effects

	TP	FP
Random guessing	100	900

Results (MM, Colombo, Kalisch & Bühlmann, Nature Methods, 2010)

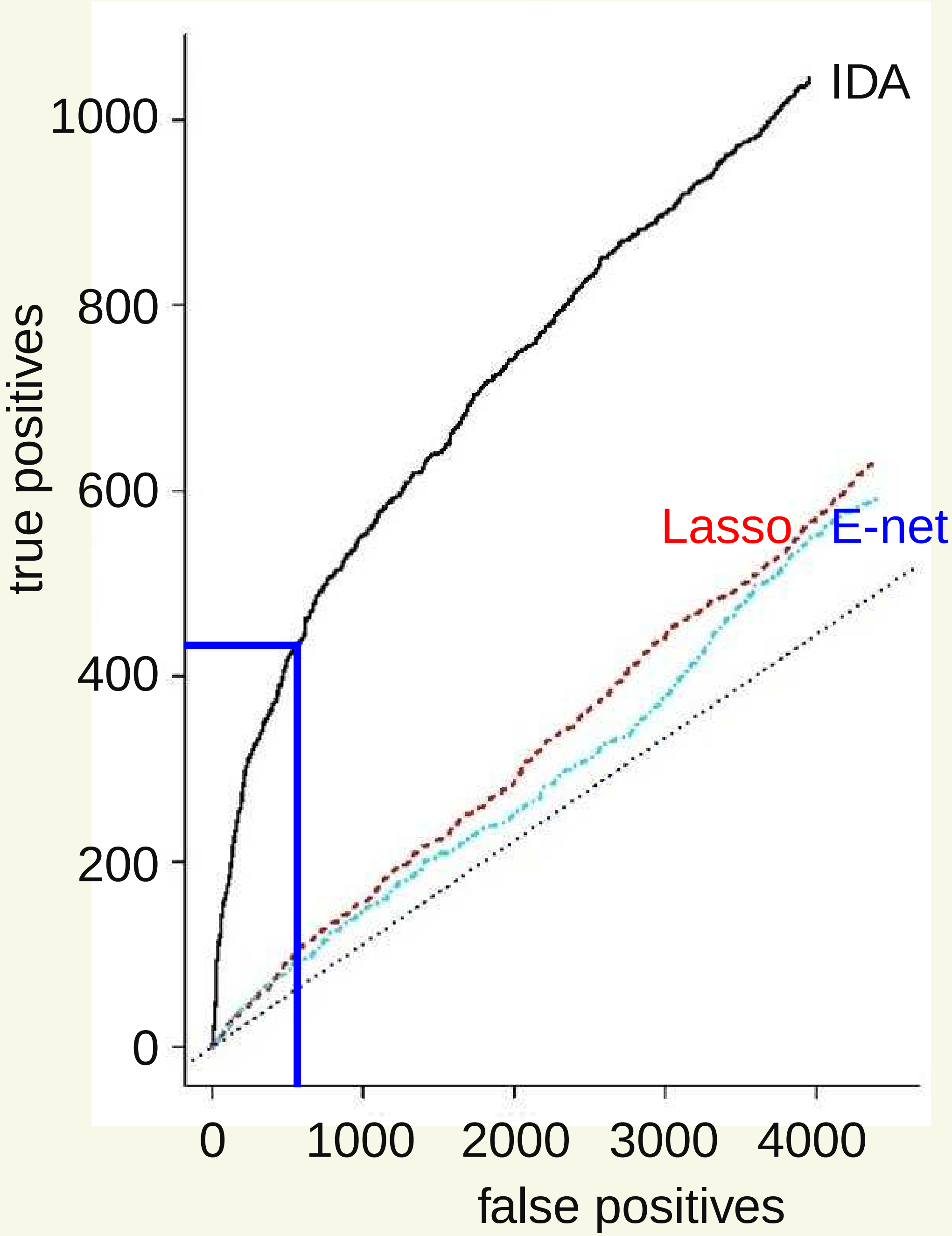


Target set: top 10% of effects from experimental data

Consider top $q = 1000$ effects

	TP	FP
Random guessing	100	900
Lasso / E-net	130	870

Results (MM, Colombo, Kalisch & Bühlmann, Nature Methods, 2010)

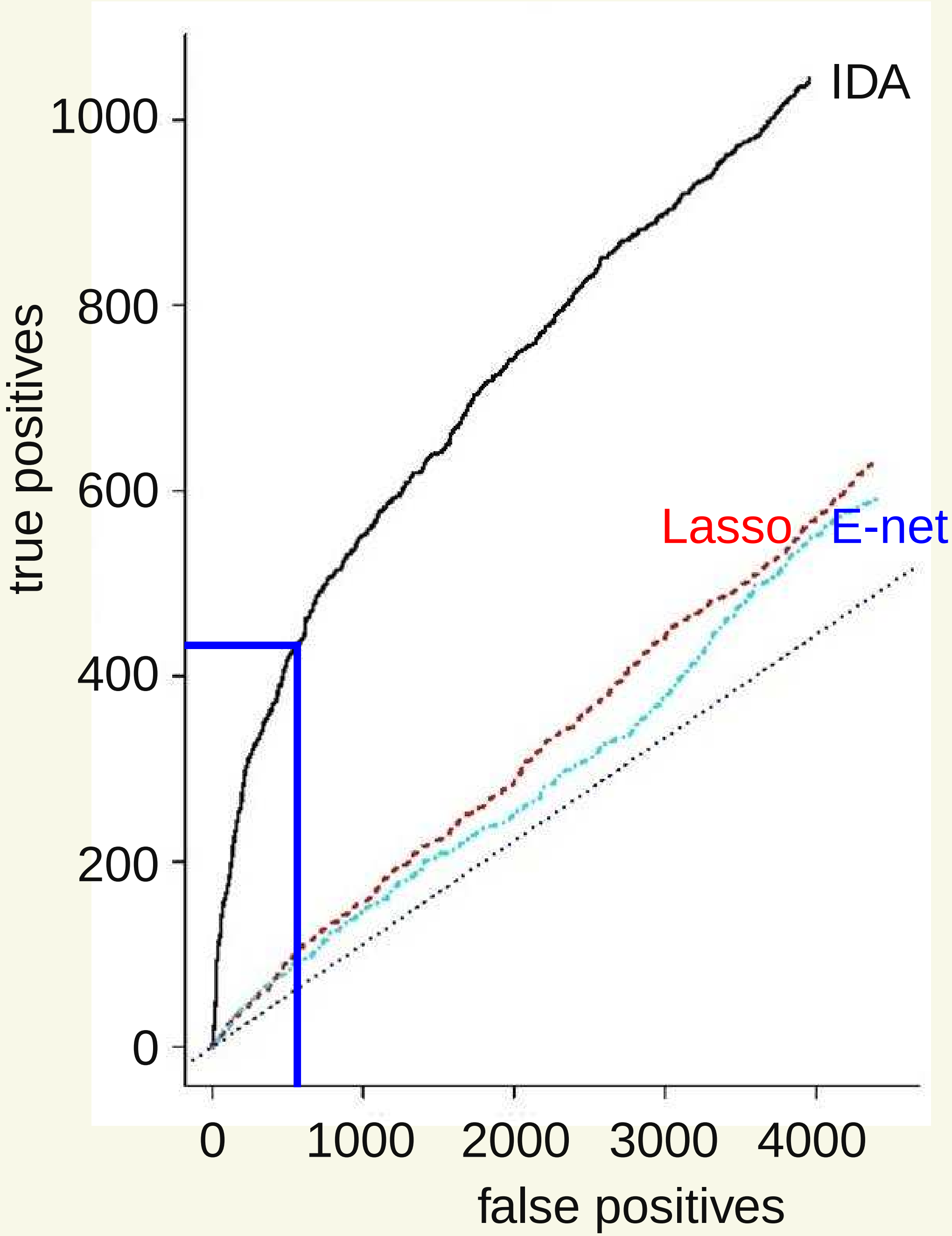


Target set: top 10% of effects from experimental data

Consider top $q = 1000$ effects

	TP	FP
Random guessing	100	900
Lasso / E-net	130	870
IDA	425	575

Results (MM, Colombo, Kalisch & Bühlmann, Nature Methods, 2010)



Target set: top 10% of effects from experimental data

Consider top $q = 1000$ effects

	TP	FP
Random guessing	100	900
Lasso / E-net	130	870
IDA	425	575

Possible use:
design of experiments

Overview

Problem: estimating causal effects from observational data in high-dimensional settings

Outline:

- What is behind IDA?
- Instability issues
- Allowing for hidden variables
- Summary and outlook

Definition of total causal effect

- Do-operator:
  - do$(X = x')$: we set the variable X to the value x' by an outside intervention, uniformly over the entire population
  - prisoner example:
    - $E(Y|do(X = 1))$ vs $E(Y|X = 1)$

Definition of total causal effect

- Do-operator:
  - do$(X = x')$: we set the variable X to the value x' by an outside intervention, uniformly over the entire population
  - prisoner example:
    - $E(Y|do(X = 1))$ vs $E(Y|X = 1)$
- Total causal effect of X on Y:
  - $\frac{\partial}{\partial x} E(Y|do(X = x))$ or
  - $E(Y|do(X = x' + 1)) - E(Y|do(X = x'))$
  - prisoner example:
    - $E(Y|do(X = 1)) - E(Y|do(X = 0))$ vs $E(Y|X = 1) - E(Y|X = 0)$

How can we estimate causal effects from observational data?

observational data

$\downarrow$

observational densities

$\downarrow$

prediction/classification:

$E(Y|X=x)$

experimental data

$\downarrow$

post-intervention densities

$\downarrow$

causal effect:

$\frac{\partial}{\partial x} E(Y|do(X=x))$

How can we estimate causal effects from observational data?

observational data

$\downarrow$

observational densities

$\downarrow$

prediction/classification:

$$E(Y|X = x)$$

experimental data

$\downarrow$

post-intervention densities

$\downarrow$

causal effect:

$$\frac{\partial}{\partial x} E(Y|do(X = x))$$

How can we estimate causal effects from observational data?

$$\begin{array}{ccc}
\text{observational data} & & \text{experimental data} \\
\downarrow & & \downarrow \\
\text{observational densities} & \dashrightarrow & \text{post-intervention densities} \\
\downarrow & & \downarrow \\
\text{prediction/classification:} & & \text{causal effect:} \\
E(Y|X = x) & & \frac{\partial}{\partial x} E(Y|do(X = x))
\end{array}$$

How can we estimate causal effects from observational data?

observational data $\downarrow$ observational densities $\downarrow$ prediction/classification: $E(Y|X=x)$

experimental data $\downarrow$ post-intervention densities $\downarrow$ causal effect: $\frac{\partial}{\partial x} E(Y|do(X=x))$

observational densities $\xrightarrow{\text{assumptions}}$ post-intervention densities

How can we estimate causal effects from observational data?

observational data

↓

observational densities $\xrightarrow{\text{assumptions}}$ post-intervention densities

↓

prediction/classification:

$E(Y|X = x)$

experimental data

↓

post-intervention densities

↓

causal effect:

$\frac{\partial}{\partial x} E(Y|do(X = x))$

Common assumption:

Causal relations are known qualitatively and can be represented by a directed acyclic graph (DAG)

Example

- Possible DAGs about sleeping problems and depression:

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

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

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

Example

- Possible DAGs about sleeping problems and depression:

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

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

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

- If the correct DAG is known, sizes of causal effects can be estimated from observational data (e.g., covariate adjustment, do-calculus, backdoor criterion, marginal structural models (Pearl, 2000; Robins et al, 1999))

Example

- Possible DAGs about sleeping problems and depression:

sleeping problems $\longrightarrow$ depression

sleeping problems $\longleftarrow$ depression

sleeping problems $\swarrow$ stress $\searrow$ depression

- If the correct DAG is known, sizes of causal effects can be estimated from observational data (e.g., covariate adjustment, do-calculus, backdoor criterion, marginal structural models (Pearl, 2000; Robins et al, 1999))
- For multivariate Gaussian data, total causal effect of X on Y is:
  - coefficient of X in the regression $Y \sim X + pa(X)$
    i.e., DAG determines adjustment variables

What can we do when the DAG is unknown?

- Knowing the DAG is unrealistic in high-dimensional settings.
  So we assume that the data come from an unknown DAG.

What can we do when the DAG is unknown?

- Knowing the DAG is unrealistic in high-dimensional settings.
  So we assume that the data come from an unknown DAG.
- A DAG encodes conditional independence relationships. So given all conditional independence relationships in the observational distribution, can we learn the DAG?

What can we do when the DAG is unknown?

- Knowing the DAG is unrealistic in high-dimensional settings.
  So we assume that the data come from an unknown DAG.
- A DAG encodes conditional independence relationships. So given all conditional independence relationships in the observational distribution, can we learn the DAG?
- Almost... several DAGs can encode the same conditional independence relationships. They are Markov equivalent.

What can we do when the DAG is unknown?

- Knowing the DAG is unrealistic in high-dimensional settings.
  So we assume that the data come from an unknown DAG.
- A DAG encodes conditional independence relationships. So given all conditional independence relationships in the observational distribution, can we learn the DAG?
- Almost... several DAGs can encode the same conditional independence relationships. They are Markov equivalent.
- Example:

	$X_1 \perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$
$X_1 \longrightarrow X_2 \longrightarrow X_3$	false	true
$X_1 \longleftarrow X_2 \longleftarrow X_3$	false	true
$X_1 \longleftarrow X_2 \longrightarrow X_3$	false	true
$X_1 \longrightarrow X_2 \longleftarrow X_3$	true	false

What can we do when the DAG is unknown?

- Knowing the DAG is unrealistic in high-dimensional settings.
  So we assume that the data come from an unknown DAG.
- A DAG encodes conditional independence relationships. So given all conditional independence relationships in the observational distribution, can we learn the DAG?
- Almost... several DAGs can encode the same conditional independence relationships. They are Markov equivalent.
- Example:

	$X_1 \perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	
$X_1 \longrightarrow X_2 \longrightarrow X_3$	false	true	
$X_1 \longleftarrow X_2 \longleftarrow X_3$	false	true	$\longleftarrow$ no v-structure
$X_1 \longleftarrow X_2 \longrightarrow X_3$	false	true	
$X_1 \longrightarrow X_2 \longleftarrow X_3$	true	false	$\longleftarrow$ v-structure

Markov equivalence class

- All DAGs in a Markov equivalence class have the same skeleton and the same v-structures

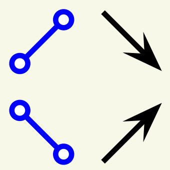
Markov equivalence class

- All DAGs in a Markov equivalence class have the same skeleton and the same v-structures
- They can be uniquely represented by a CPDAG:
  - edge between X and Y iff $X \not\perp\!\!\!\perp Y | \mathbf{S}$ for all subsets $\mathbf{S}$ of the remaining variables
    (edges are stronger than in CIGs/Gaussian models)
  - $X \rightarrow Y$ iff $X \rightarrow Y$ in all DAGs in the equivalence class
    (direct causal effect)
  - $X \circ\!\!-\!\!\circ Y$ iff there is a DAG in the equivalence class with $X \rightarrow Y$ and one with $X \leftarrow Y$
    (unidentifiable orientations)

Markov equivalence class

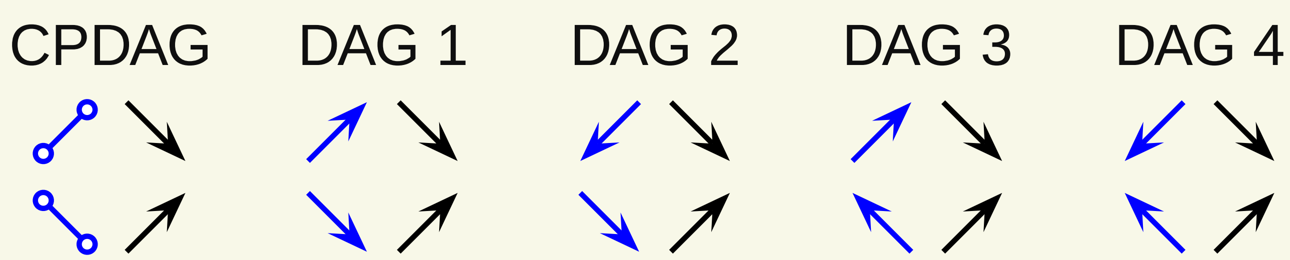
- All DAGs in a Markov equivalence class have the same skeleton and the same v-structures
- They can be uniquely represented by a CPDAG:
    - edge between X and Y iff $X \not\perp\!\!\!\perp Y|\mathbf{S}$ for all subsets $\mathbf{S}$ of the remaining variables
      (edges are stronger than in CIGs/Gaussian models)
    - $X \to Y$ iff $X \to Y$ in all DAGs in the equivalence class
      (direct causal effect)
    - $X \circ\!\!-\!\!\circ Y$ iff there is a DAG in the equivalence class with $X \to Y$ and one with $X \leftarrow Y$
      (unidentifiable orientations)
- Example:

CPDAG



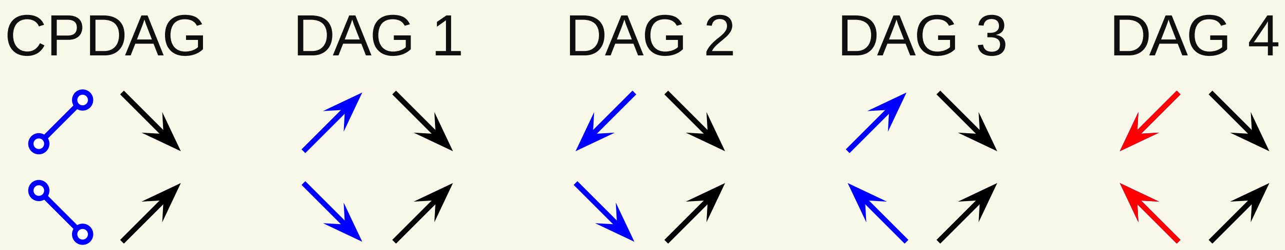
Markov equivalence class

- All DAGs in a Markov equivalence class have the same skeleton and the same v-structures
- They can be uniquely represented by a CPDAG:
  - edge between X and Y iff $X \not\perp\!\!\!\perp Y|\mathbf{S}$ for all subsets $\mathbf{S}$ of the remaining variables
    (edges are stronger than in CIGs/Gaussian models)
  - $X \rightarrow Y$ iff $X \rightarrow Y$ in all DAGs in the equivalence class
    (direct causal effect)
  - $X \circ\!\!-\!\!\circ Y$ iff there is a DAG in the equivalence class with $X \rightarrow Y$ and one with $X \leftarrow Y$
    (unidentifiable orientations)
- Example:



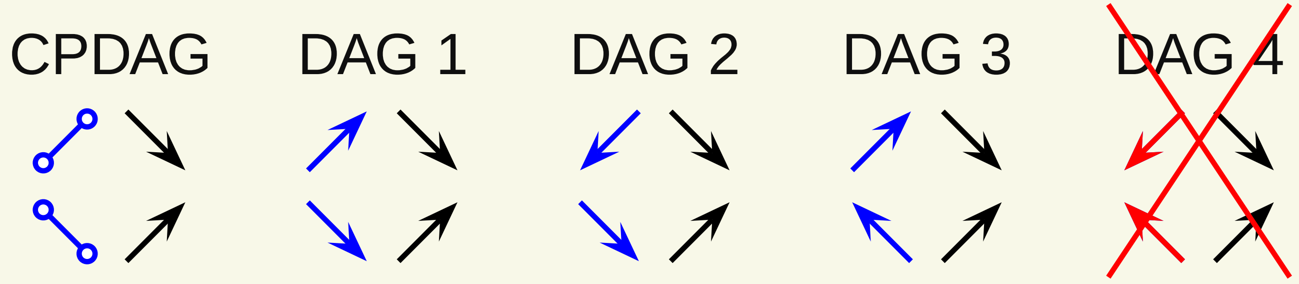
Markov equivalence class

- All DAGs in a Markov equivalence class have the same skeleton and the same v-structures
- They can be uniquely represented by a CPDAG:
  - edge between X and Y iff $X \not\perp\!\!\!\perp Y | \mathbf{S}$ for all subsets $\mathbf{S}$ of the remaining variables
    (edges are stronger than in CIGs/Gaussian models)
  - $X \rightarrow Y$ iff $X \rightarrow Y$ in all DAGs in the equivalence class
    (direct causal effect)
  - $X \circ\!\!-\!\!\circ Y$ iff there is a DAG in the equivalence class with $X \rightarrow Y$ and one with $X \leftarrow Y$
    (unidentifiable orientations)
- Example:



Markov equivalence class

- All DAGs in a Markov equivalence class have the same skeleton and the same v-structures
- They can be uniquely represented by a CPDAG:
  - edge between X and Y iff $X \not\perp\!\!\!\perp Y | \mathbf{S}$ for all subsets $\mathbf{S}$ of the remaining variables
    (edges are stronger than in CIGs/Gaussian models)
  - $X \rightarrow Y$ iff $X \rightarrow Y$ in all DAGs in the equivalence class
    (direct causal effect)
  - $X \circ\!\!-\!\!\circ Y$ iff there is a DAG in the equivalence class with $X \rightarrow Y$ and one with $X \leftarrow Y$
    (unidentifiable orientations)
- Example:



Causal structure learning

- Learning CPDAGs is challenging. Main methods:
  - Score-based methods: e.g. Penalized MLE, GES (Chickering, 2002; Peter's talk, Alain's talk)
  - Constraint-based methods: e.g. PC algorithm (Spirtes et al, 2000)

Causal structure learning

- Learning CPDAGs is challenging. Main methods:
  - Score-based methods: e.g. Penalized MLE, GES (Chickering, 2002; Peter’s talk, Alain’s talk)
  - Constraint-based methods: e.g. PC algorithm (Spirtes et al, 2000)
    - Fast
    - Requires faithfulness assumption: the conditional independencies in the distribution equal the ones encoded in the DAG
    - Consistent for high-dimensional sparse graphs (Kalisch and Bühlmann, JMLR, 2007)

Causal structure learning

- Learning CPDAGs is challenging. Main methods:
  - Score-based methods: e.g. Penalized MLE, GES (Chickering, 2002; Peter's talk, Alain's talk)
  - Constraint-based methods: e.g. PC algorithm (Spirtes et al, 2000)
    - Fast
    - Requires faithfulness assumption: the conditional independencies in the distribution equal the ones encoded in the DAG
    - Consistent for high-dimensional sparse graphs (Kalisch and Bühlmann, JMLR, 2007)
  - Restricted structural equation models: e.g. LiNGAM (Shimizu et al, 2006; Tübingen group; Jonas's talk)

Estimating causal effects when equivalence class is given

- Due to equivalence class, the problem of estimating causal effects is under-determined. Hence, effects may be unidentifiable.

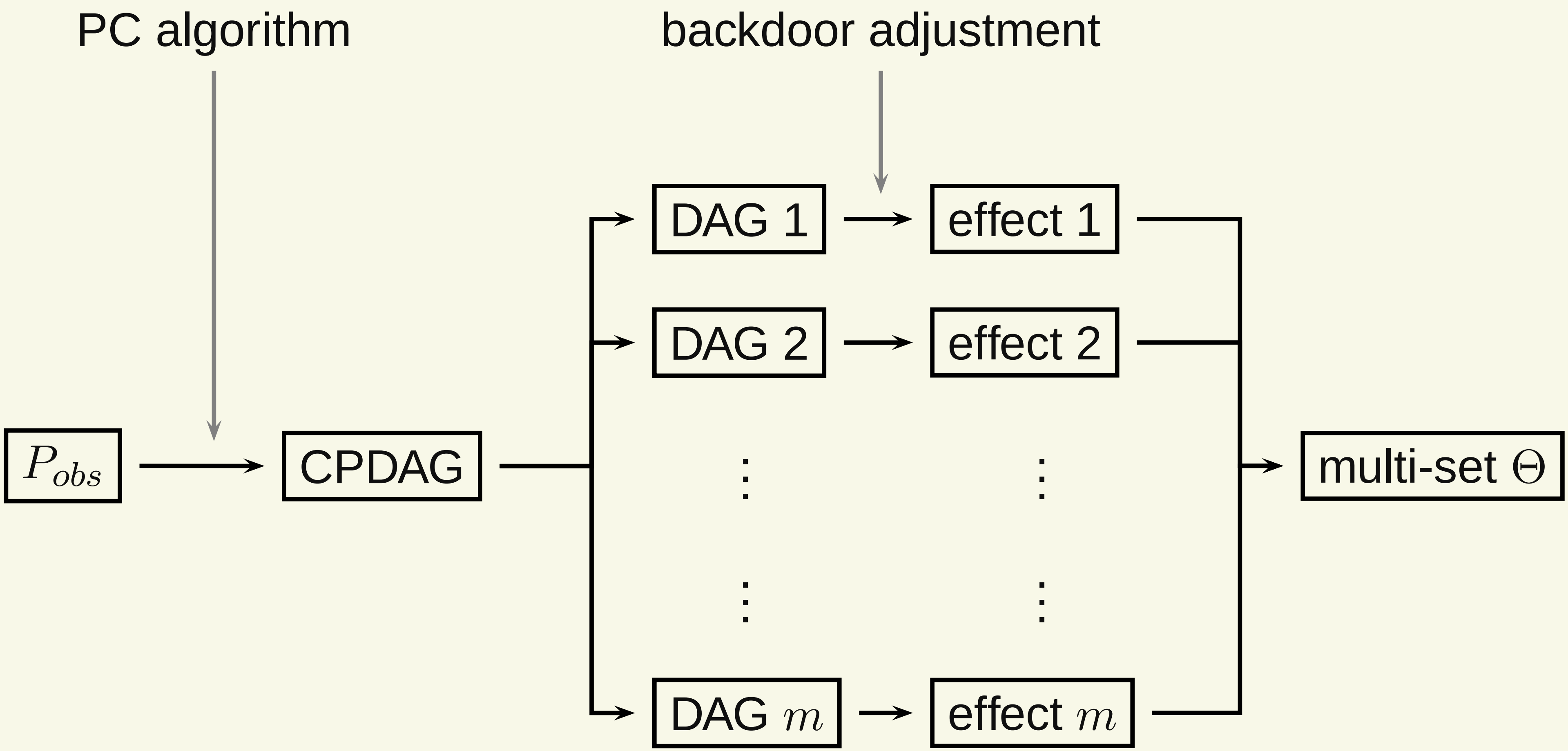
Estimating causal effects when equivalence class is given

- Due to equivalence class, the problem of estimating causal effects is under-determined. Hence, effects may be unidentifiable.
- Existing work: Prediction algorithm (Spirtes et al, 2000)
  - Estimates only identifiable causal effects
  - Goes through all possible orderings of the variables; computationally infeasible for large problems

Estimating causal effects when equivalence class is given

- Due to equivalence class, the problem of estimating causal effects is under-determined. Hence, effects may be unidentifiable.
- Existing work: Prediction algorithm (Spirtes et al, 2000)
  - Estimates only identifiable causal effects
  - Goes through all possible orderings of the variables; computationally infeasible for large problems
- IDA algorithm (MM, Kalisch, Bühlmann, AoS, 2009)
  - Intervention-calculus when the DAG is Absent
  - Estimates multi-sets of possible causal effects
  - Local method that scales well to large graphs

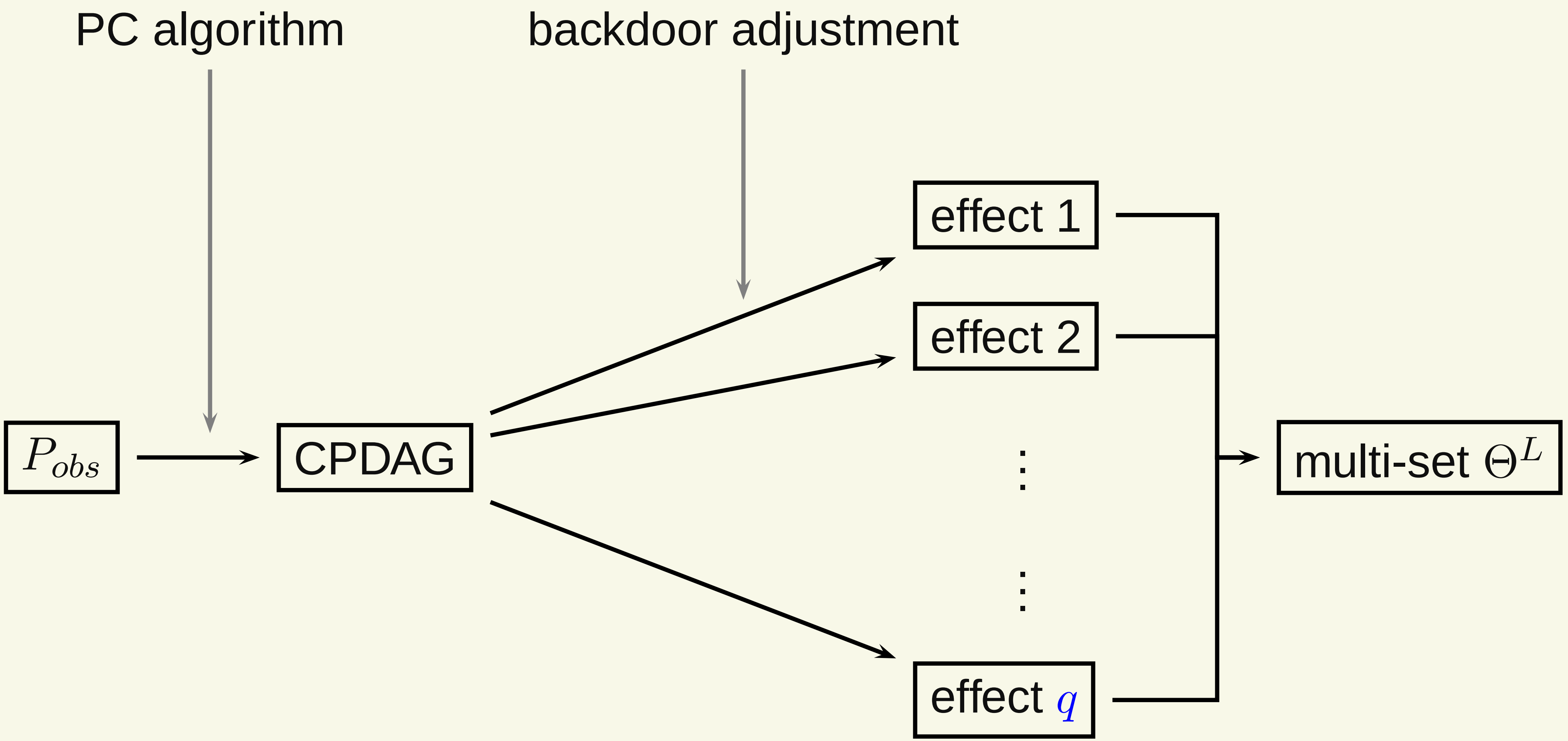
IDA (oracle version)



The true causal effect is in Θ.

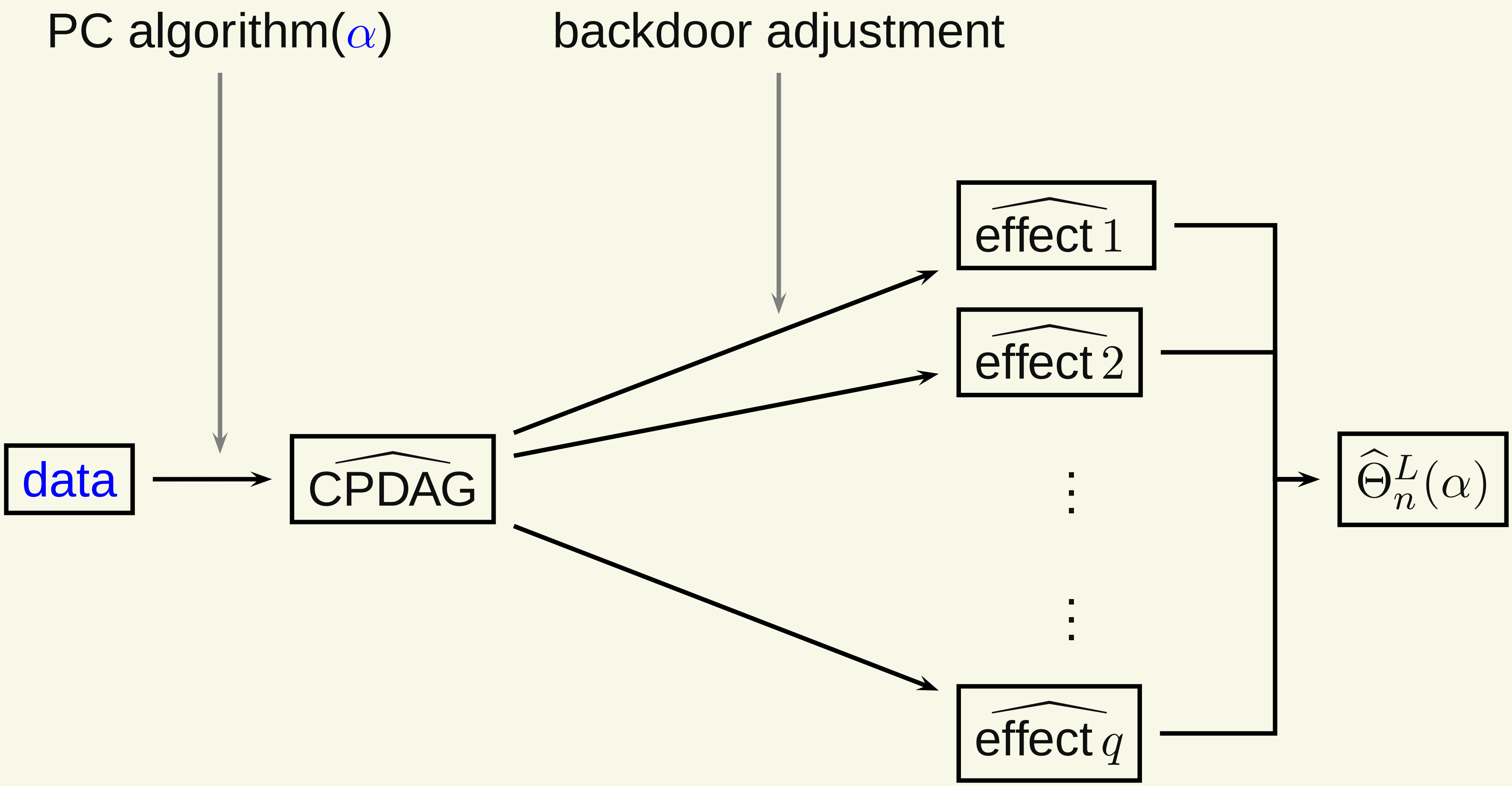
We can obtain bounds for the size of the causal effect.

IDA (local oracle version)



Bounds based on Θ^L are identical to bounds based on Θ.
Proof uses graph theoretic properties of the CPDAG.

IDA (local sample version)



The estimates are consistent in sparse high-dimensional settings

Summary of IDA

- IDA estimates bounds on causal effects from observational data, assuming the data come from an unknown DAG:
  - computationally feasible for large sparse systems due to PC algorithm and local method
  - software: R-package `pcalg`
    (Kalisch et al, JSS, 2012)
  - consistency in sparse high-dimensional settings
    (MM, Kalisch and Bühlmann, AoS, 2009)
  - validations in biological systems
    (MM, Colombo, Kalisch and Bühlmann, Nature Methods, 2010; Stekhoven et al, Bioinformatics, 2012)

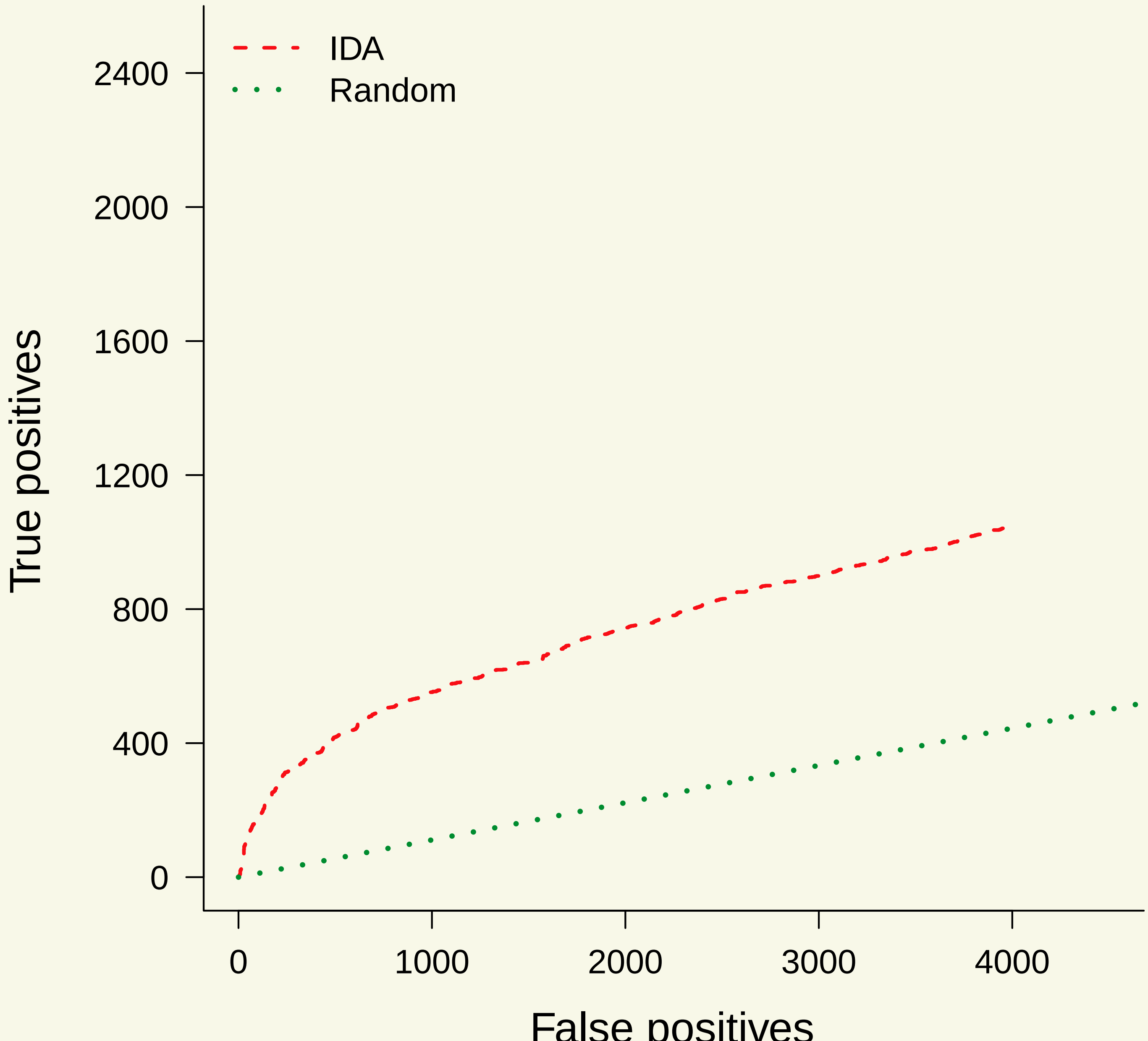
Overview

Problem: estimating causal effects from observational data in high-dimensional settings

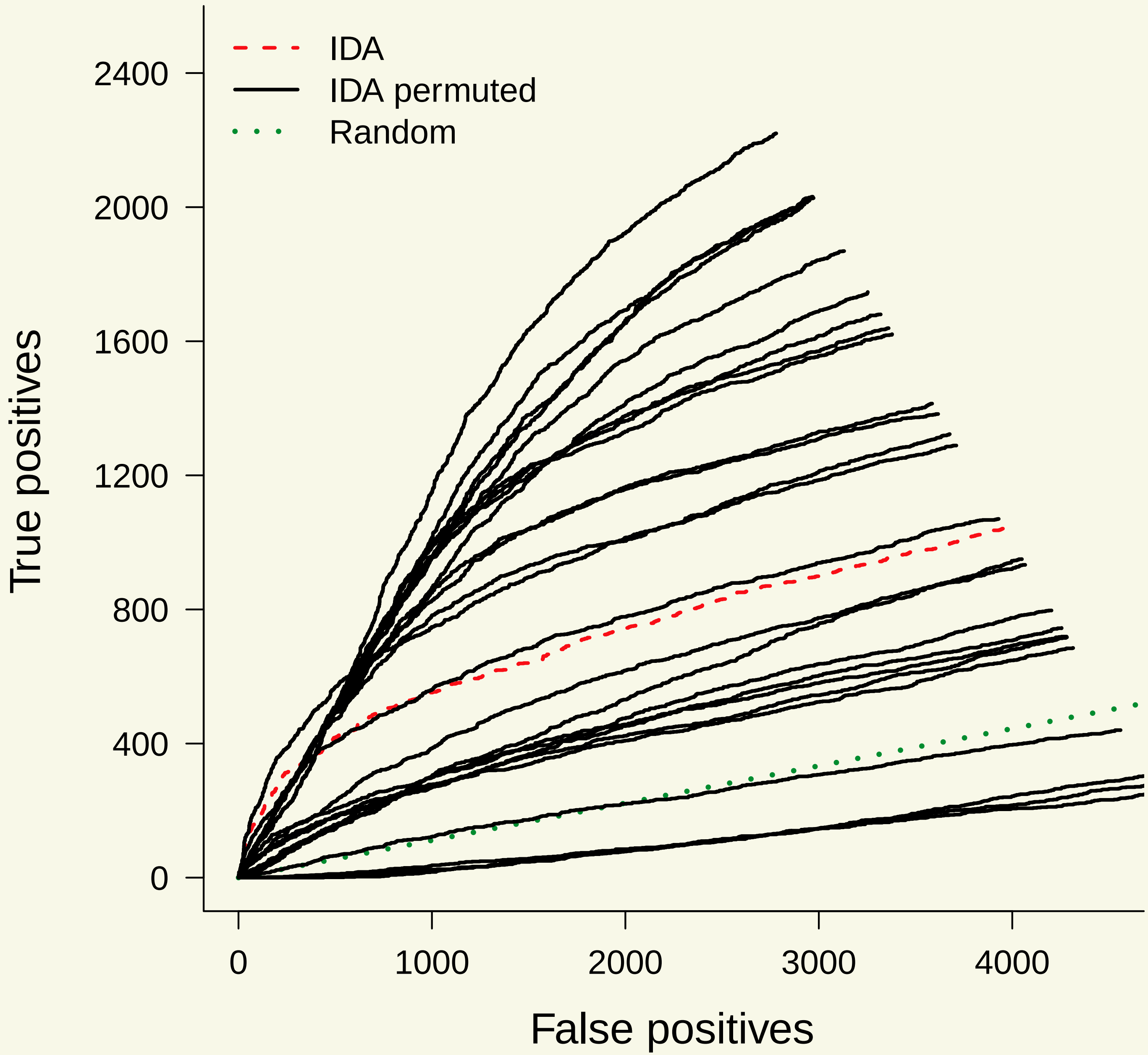
Outline:

- What is behind IDA?
- Instability issues
- Allowing for hidden variables
- Summary and outlook

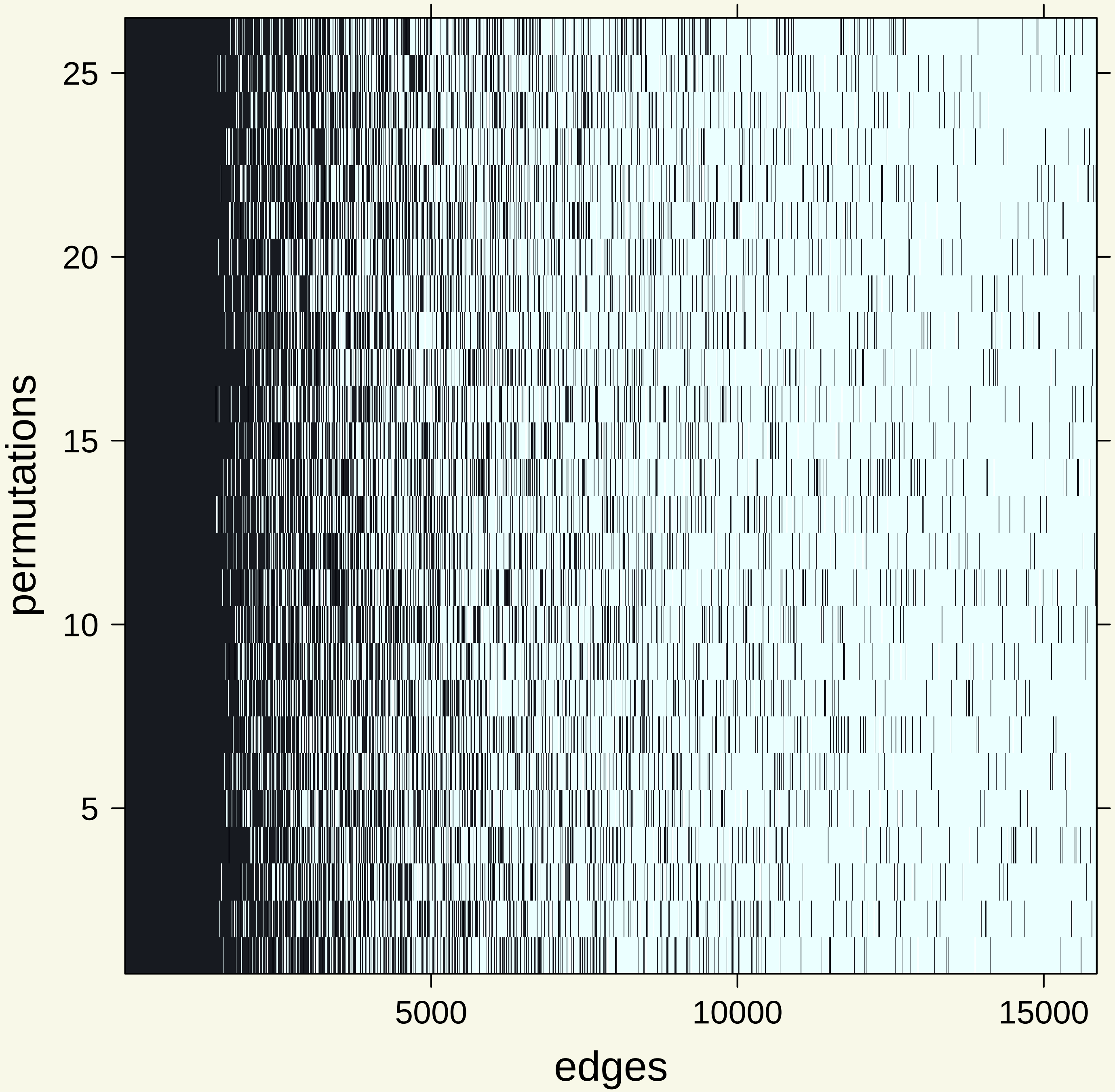
Yeast: IDA results



Yeast: IDA results highly sensitive to variable ordering



Yeast: order-dependence in PC algorithm



PC algorithm: three steps

- Determine the skeleton
- Determine the v-structures
- Orient as many of the remaining edges as possible

PC algorithm: the skeleton

- Idea:
  - No edge between X_i and X_j
    $\iff$
    $X_i \perp\!\!\!\perp X_j | S$ for some subset S of the remaining variables
    $\iff$
    $X_i \perp\!\!\!\perp X_j | S'$ for some subset S' of adj(X_i) or adj(X_j)

PC algorithm: the skeleton

- Idea:
  - No edge between X_i and X_j
    $\iff$
    $X_i \perp\!\!\!\perp X_j | S$ for some subset S of the remaining variables
    $\iff$
    $X_i \perp\!\!\!\perp X_j | S'$ for some subset S' of adj(X_i) or adj(X_j)
- Implementation:
  - Start with complete graph
  - Consider marginal independence of all pairs of variables; remove edges
  - Consider conditional independence given subsets of size k of the adjacency sets, for $k = 1, 2, \ldots$

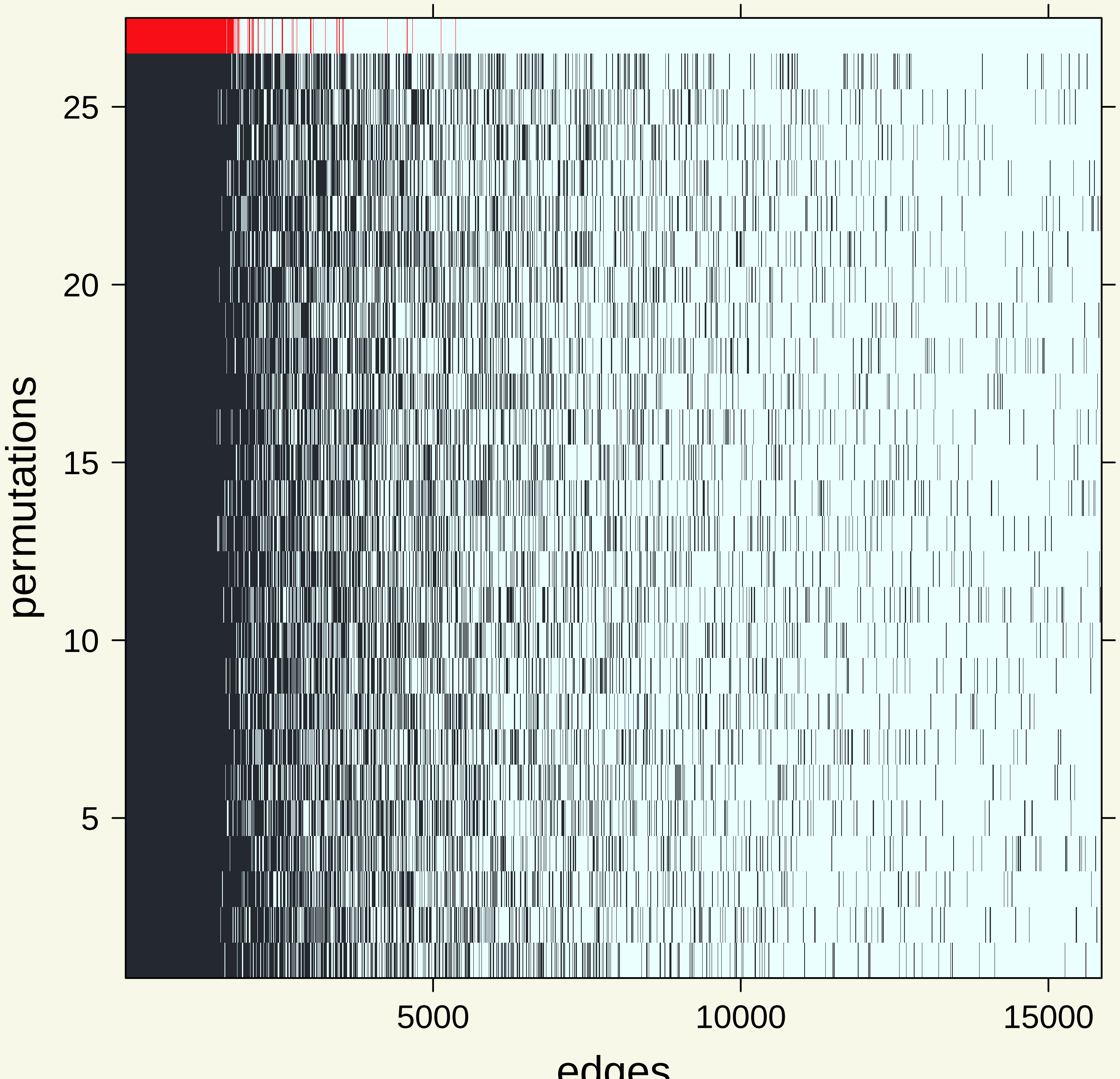
PC algorithm: the skeleton

- Idea:
  - No edge between X_i and X_j
    $\Longleftrightarrow$
    $X_i \perp\!\!\!\perp X_j | S$ for some subset S of the remaining variables
    $\Longleftrightarrow$
    $X_i \perp\!\!\!\perp X_j | S'$ for some subset S' of adj(X_i) or adj(X_j)
- Implementation:
  - Start with complete graph
  - Consider marginal independence of all pairs of variables; remove edges
  - Consider conditional independence given subsets of size k of the adjacency sets, for $k = 1, 2, \ldots$
- Order-dependence in sample version: the order in which variables are tested determines which edges are removed first. This affects which tests are considered later on.

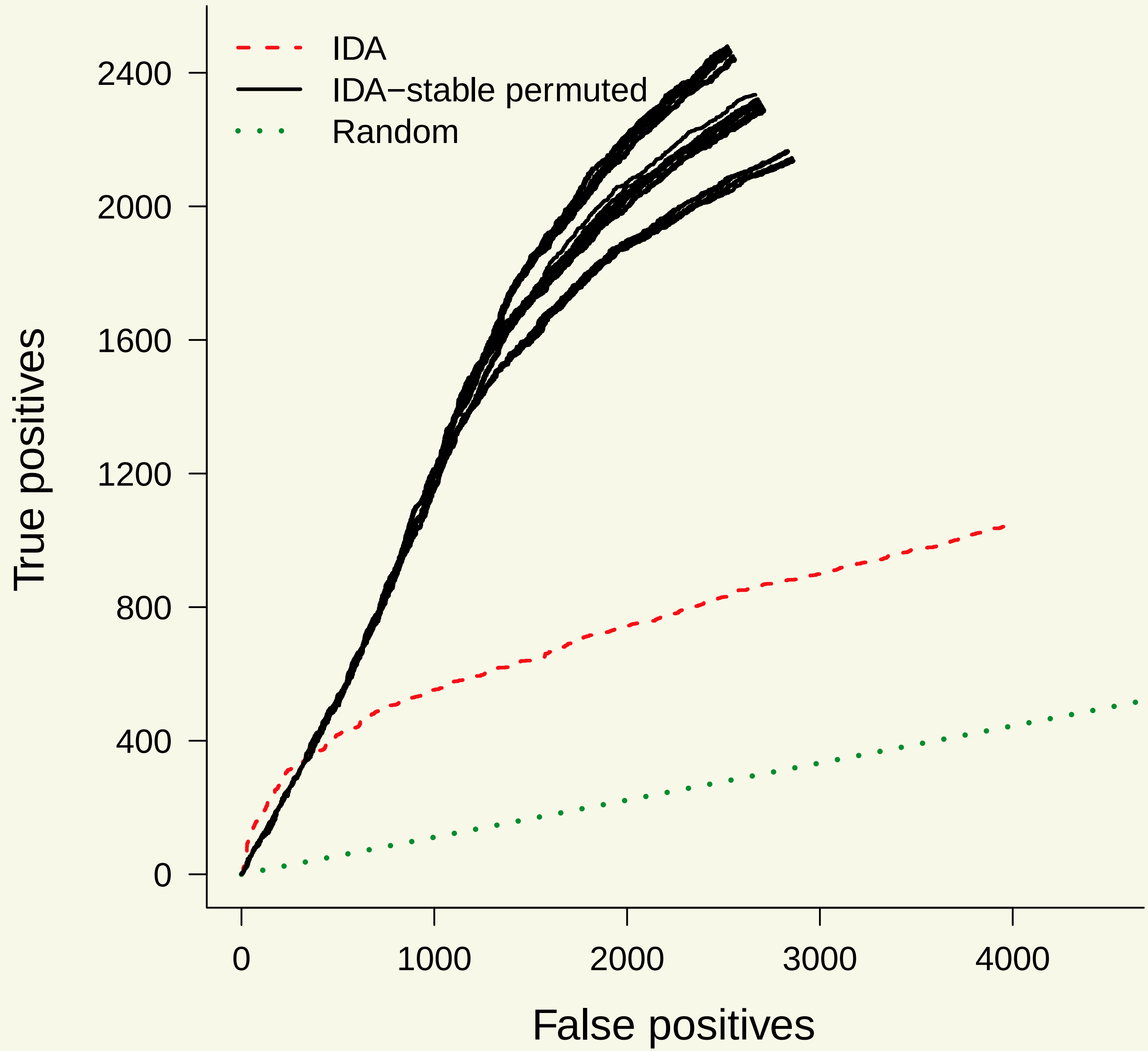
PC algorithm: the skeleton

- Idea:
  - No edge between X_i and X_j
    $\iff$
    $X_i \perp\!\!\!\perp X_j | S$ for some subset S of the remaining variables
    $\iff$
    $X_i \perp\!\!\!\perp X_j | S'$ for some subset S' of adj(X_i) or adj(X_j)
- Implementation:
  - Start with complete graph
  - Consider marginal independence of all pairs of variables; remove edges
  - Consider conditional independence given subsets of size k of the adjacency sets, for $k = 1, 2, \ldots$
- Order-dependence in sample version: the order in which variables are tested determines which edges are removed first. This affects which tests are considered later on.
- PC-stable: don't change skeleton within each level of k

Yeast: PC-stable algorithm (Colombo and MM, arXiv 2012)



Yeast: IDA-stable results



Order-dependence: main points

- Order-dependence of the PC algorithm can lead to highly variable results in high-dimensional settings

Order-dependence: main points

- Order-dependence of the PC algorithm can lead to highly variable results in high-dimensional settings
- PC-stable is a simple modification of the PC algorithm:
  - oracle version is identical to PC
  - sample version:
    - yields order-independent skeletons
    - is consistent under the same conditions as PC
    - “automatically” picks out the stable edges in high-dimensional sample version

Order-dependence: main points

- Order-dependence of the PC algorithm can lead to highly variable results in high-dimensional settings
- PC-stable is a simple modification of the PC algorithm:
    - oracle version is identical to PC
    - sample version:
        - yields order-independent skeletons
        - is consistent under the same conditions as PC
        - “automatically” picks out the stable edges in high-dimensional sample version
- IDA based on PC-stable can lead to large improvements
- Results can be further improved by incorporating subsampling (Stekhoven et al, Bioinformatics, 2012)

Overview

Problem: estimating causal effects from observational data in high-dimensional settings

Outline:

- What is behind IDA?
- Instability issues
- Allowing for hidden variables
- Summary and outlook

Example: what can go wrong when there are hidden variables?

- DAG with two hidden variables L_1 and L_2:

$$\begin{tikzcd}
 & L_1 \arrow[dl] \arrow[dr] & & L_2 \arrow[dl] \arrow[dr] & \\
X_1 & & X_2 & & X_3
\end{tikzcd}$$

Example: what can go wrong when there are hidden variables?

- DAG with two hidden variables L_1 and L_2:

$$\begin{tikzcd}
 & L_1 \arrow[dl] \arrow[dr] & & L_2 \arrow[dl] \arrow[dr] & \\
X_1 & & X_2 & & X_3
\end{tikzcd}$$

- The conditional independence relationships among $\{X_1, X_2, X_3\}$ can be uniquely represented by the DAG

$$\begin{tikzcd}
X_1 \arrow[r] & X_2 & X_3 \arrow[l]
\end{tikzcd}$$

Example: what can go wrong when there are hidden variables?

- DAG with two hidden variables L_1 and L_2:

$$\begin{tikzcd} & L_1 \arrow[dl] \arrow[dr] & & L_2 \arrow[dl] \arrow[dr] & \\ X_1 & & X_2 & & X_3 \end{tikzcd}$$

- The conditional independence relationships among $\{X_1, X_2, X_3\}$ can be uniquely represented by the DAG

$$X_1 \longrightarrow X_2 \longleftarrow X_3$$

- But interpreting this DAG causally would lead us to think that X_1 and X_3 are causes of X_2.
- This is wrong! And this causes the output of IDA to be wrong!

Approach in the presence of hidden variables

- We work with maximal ancestral graphs (MAGs) (Richardson and Spirtes, 2002)
- We consider the following two steps:
  - Learning the equivalence class of MAGs
  - Estimating causal effects when the equivalence class is given

Approach in the presence of hidden variables

- We work with maximal ancestral graphs (MAGs) (Richardson and Spirtes, 2002)
- We consider the following two steps:
  - Learning the equivalence class of MAGs
  - Estimating causal effects when the equivalence class is given

- Note:
  - Since we allow hidden variables, the equivalence classes are larger, and the problem is even more underdetermined.
  - But we can still learn causal information. In the example, conditional independence relationships among $\{X_1, X_2, X_3\}$ imply:
    - X_2 is not a cause of X_1 nor of X_3
    - X_1 is not a cause of X_3 and vice versa
    - X_1 and X_3 may or may not be causes of X_2

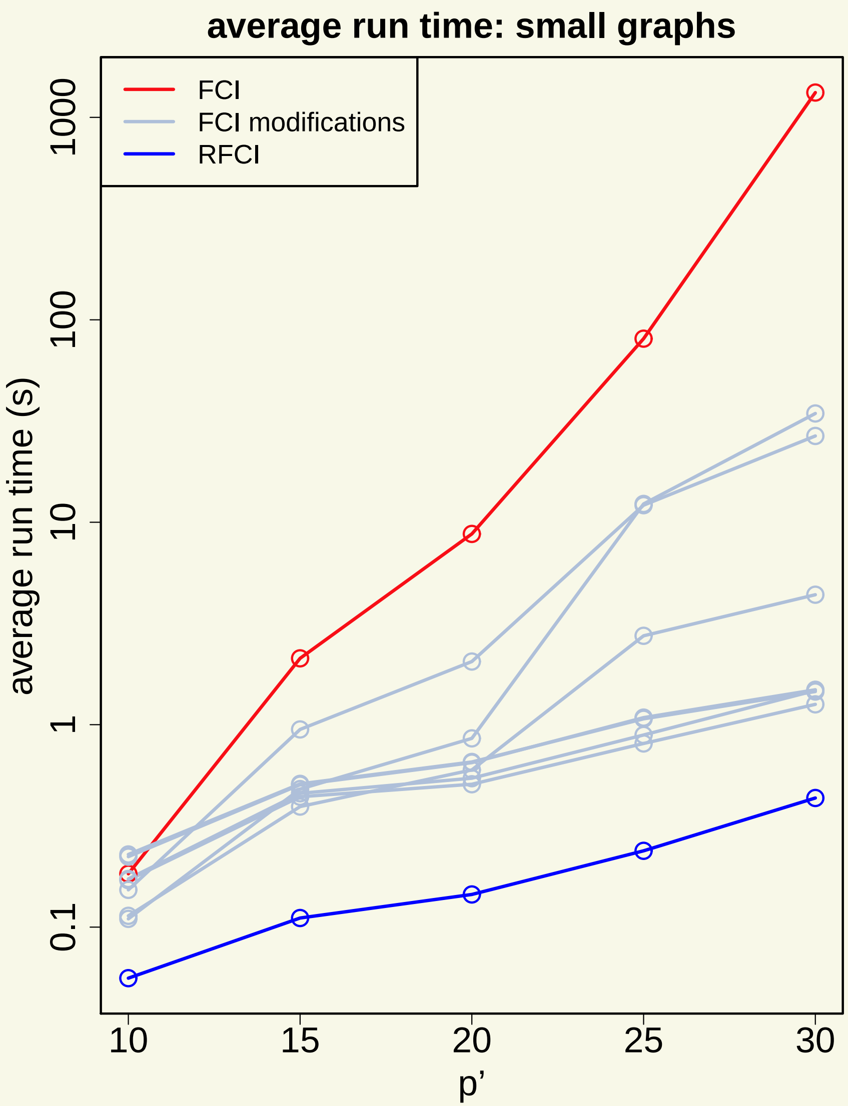
Causal structure learning

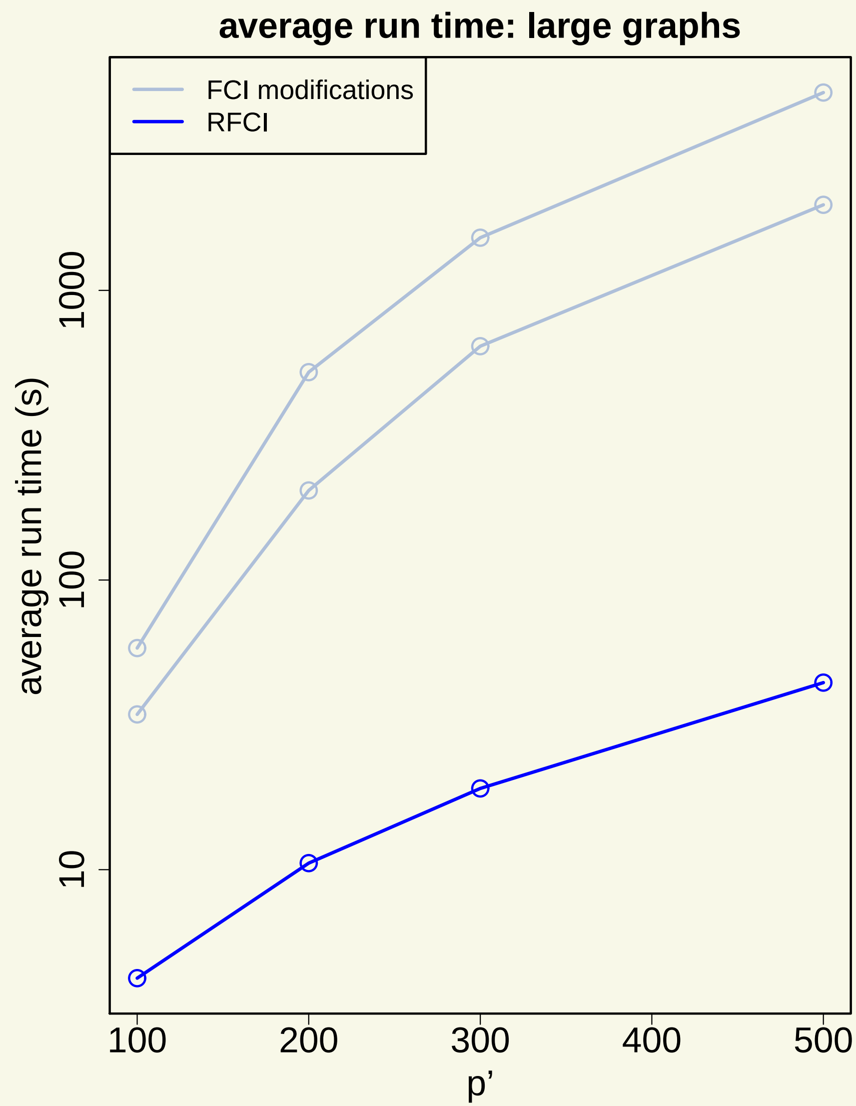
- Existing work: FCI algorithm (Spirtes et al, 1999)

Causal structure learning

- Existing work: FCI algorithm (Spirtes et al, 1999)
- Our work (Colombo, MM, Kalisch and Richardson, 2012):
  - Fast modifications of FCI:
    - Perform fewer tests
    - Identical in the oracle version
  - RFCI algorithm:
    - Performs only local tests; speed comparable to PC
    - Output slightly less informative than FCI
    - Correct causal interpretation
    - Sufficient conditions on underlying DAG for equality of FCI and RFCI
  - Consistency in high-dimensional settings for all algorithms (under weaker conditions for RFCI)

Causal structure learning: computational performance

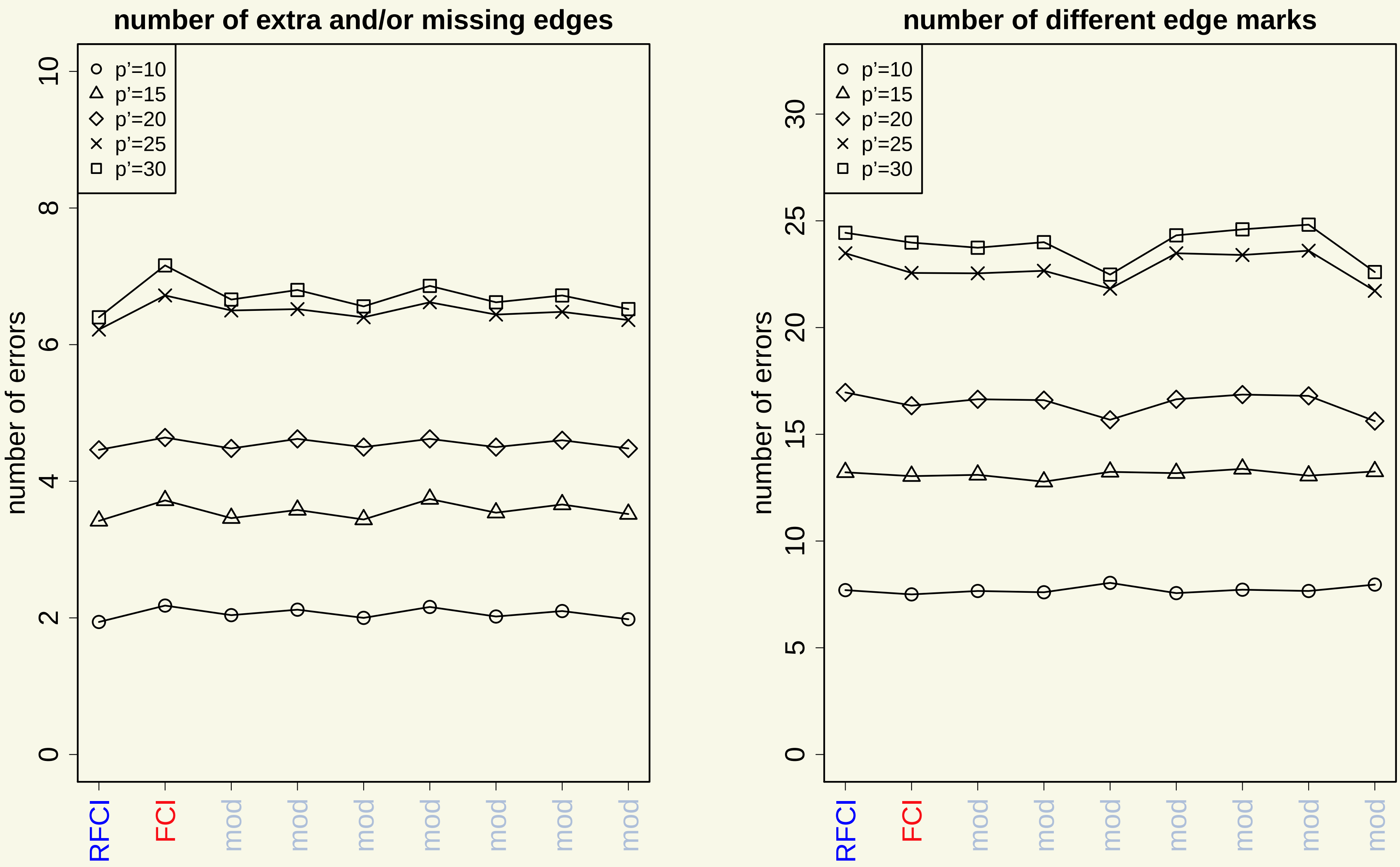




Modifications of FCI achieve a large speed-up.

RFCI is the only algorithm that scales well to large graphs.

Causal structure learning: estimation performance



All algorithms have comparable estimation performance.

Summary

- Problem:
  - There is a need for causal methods for observational data from complex systems
  - Such methods cannot replace randomized controlled experiments. But they can be very useful for exploration:
    - hypothesis generation
    - prioritization of experiments

Summary

- IDA estimates bounds on causal effects from observational data, assuming the data come from an unknown DAG:
  - computationally feasible for large sparse systems
  - consistency in sparse high-dimensional settings
  - validations in biological systems

Summary

- IDA estimates bounds on causal effects from observational data, assuming the data come from an unknown DAG:
  - computationally feasible for large sparse systems
  - consistency in sparse high-dimensional settings
  - validations in biological systems
- Beware of possible instability for high-dimensional data
  - instability due to order-dependence is mostly resolved

Summary

- IDA estimates bounds on causal effects from observational data, assuming the data come from an unknown DAG:
  - computationally feasible for large sparse systems
  - consistency in sparse high-dimensional settings
  - validations in biological systems
- Beware of possible instability for high-dimensional data
  - instability due to order-dependence is mostly resolved
- In the presence of unmeasured variables:
  - RFCI and modifications of FCI for fast causal structure learning
  - consistency in sparse high-dimensional settings

Summary

- IDA estimates bounds on causal effects from observational data, assuming the data come from an unknown DAG:
  - computationally feasible for large sparse systems
  - consistency in sparse high-dimensional settings
  - validations in biological systems
- Beware of possible instability for high-dimensional data
  - instability due to order-dependence is mostly resolved
- In the presence of unmeasured variables:
  - RFCI and modifications of FCI for fast causal structure learning
  - consistency in sparse high-dimensional settings
- All software is available in the R-package `pcalg` (Kalisch et al, JSS, 2012)

Summary

- Some current/future work:
  - causal structure learning (talks of Peter, Jonas and Alain):
    - alternatives to the PC algorithm / faithfulness assumption
    - allowing for non-Gaussian data / mixed variables
    - allowing for heterogeneous data / mix of observational and interventional data
    - incorporating time series data

Summary

- Some current/future work:
  - causal structure learning (talks of Peter, Jonas and Alain):
    - alternatives to the PC algorithm / faithfulness assumption
    - allowing for non-Gaussian data / mixed variables
    - allowing for heterogeneous data / mix of observational and interventional data
    - incorporating time series data
  - principled methods for choice of tuning parameter
  - deriving more statistical properties (i.e., standard errors)
  - generalizing to multiple simultaneous interventions
  - IDA that allows for hidden variables
  - more applications and validations on interesting data sets

Thank you for your attention!

References

- Colombo, Maathuis, Kalisch and Richardson (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294-321.
- Colombo and Maathuis (2012). A modification of the PC algorithm yielding order-independent skeletons. *Submitted* (arXiv:1211.3295)
- Hughes et al (2000). Functional discovery via a compendium of expression profiles. *Cell* **102** 109-126.
- Kalisch, Mächler, Colombo, Maathuis and Bühlmann (2012). Causal inference using graphical models with the R-package pcalg. *J. Statist. Software* **47** 11 1-26.
- Maathuis, Colombo, Kalisch and Bühlmann (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods* **7** 247-248.
- Maathuis, Kalisch and Bühlmann (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133-3164.
- Meinshausen and Bühlmann (2010). Stability Selection. *JRSS B* **72** 417-473.
- Pearl (2000). *Causality.* Cambridge University Press.
- Spirtes, Glymour and Scheines (2000). *Causation, Prediction and Search.* MIT Press.
- Stekhoven, Hennig, Sveinbjörnsson, Moraes, Maathuis and Bühlmann (2012). Causal stability ranking. *Bioinformatics* **28** 2819-2823.