Inhomogenous large-scale data:
new opportunities for causal inference and prediction

Peter Bühlmann

Seminar für Statistik, ETH Zürich

joint work with

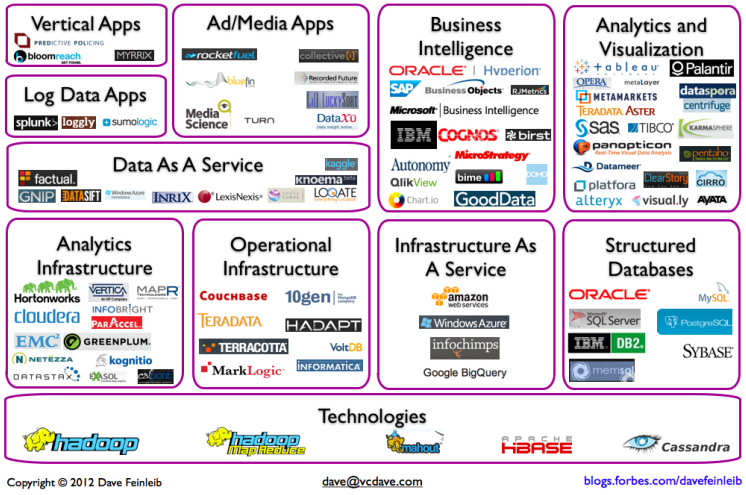
Jonas Peters
Univ. Copenhagen

Nicolai Meinshausen
SfS ETH Zürich

Heterogeneous large-scale data

Big Data

Big Data Landscape



the talk is not (yet) on “really big data”

but we will take advantage of heterogeneity

often arising with large-scale data where

i.i.d./homogeneity assumption is not appropriate

Two seemingly different problems

1. prediction in heterogeneous environments
2. causal inference = intervention analysis

but they are very closely related!

Two seemingly different problems

1. prediction in heterogeneous environments
2. causal inference = intervention analysis

but they are very closely related!

1. Prediction in heterogeneous environments

data from different known observed environments/experimental conditions/sub-populations $e \in \mathcal{E}$:

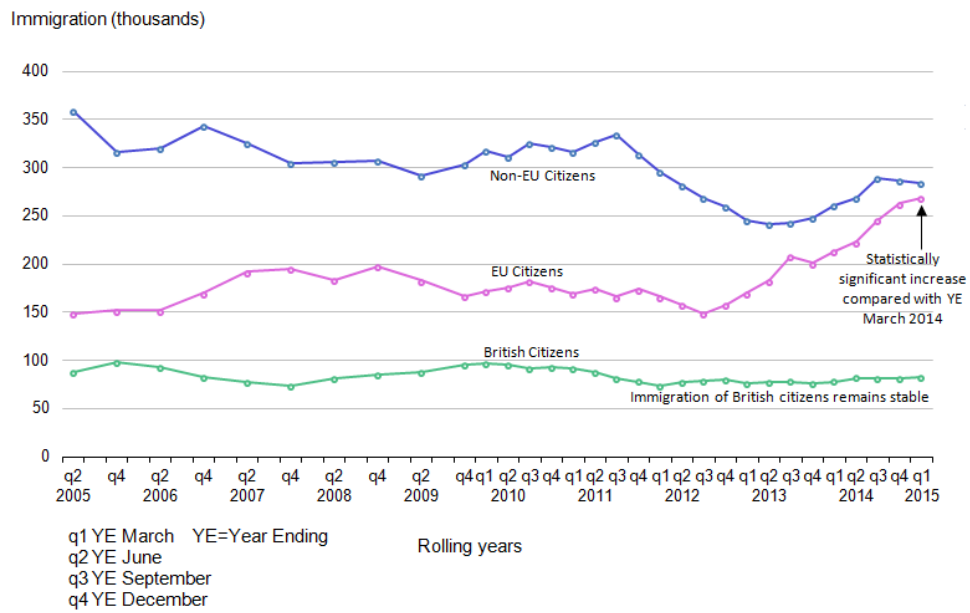
$$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

with response variables Y^e and predictor variables X^e

examples:

- data from 10 different countries
- data from economic scenarios (from different "time blocks")

immigration in the UK

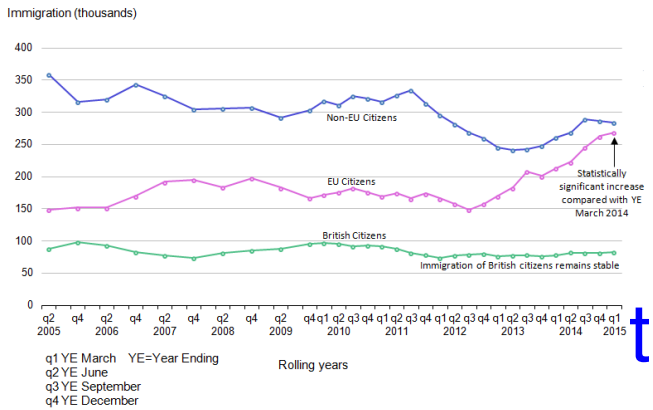


consider “all possible” but
mostly non-observed environments $\mathcal{F} \supset \underbrace{\mathcal{E}}_{\text{observed}}$

examples for $\mathcal{F}$:

- 10 countries and many other than the 10 countries
- the presence and the unseen future with new scenarios

immigration in the UK



the unseen future

problem:

predict Y given X such that the prediction works well (is “robust”) *for “all possible”* environments $e \in \mathcal{F}$
based on data from much fewer environments from $\mathcal{E}$

problem:
predict Y given X such that the prediction works well (is "robust") *for "all possible"* environments $e \in \mathcal{F}$
based on data from much fewer environments from $\mathcal{E}$

we need a model, of course! (one which is good/"justifiable")

and we will illustrate "validated" examples from genomics
with respect to predicting values in new unseen environments

problem:
predict Y given X such that the prediction works well (is "robust") *for "all possible"* environments $e \in \mathcal{F}$
based on data from much fewer environments from $\mathcal{E}$

we need a model, of course! (one which is good/"justifiable")

and we will illustrate "validated" examples from genomics
with respect to predicting values in new unseen environments

problem:
predict Y given X such that the prediction works well (is "robust") *for "all possible"* environments $e \in \mathcal{F}$
based on data from much fewer environments from $\mathcal{E}$

we need a model, of course! (one which is good/"justifiable")

and we will illustrate "validated" examples from genomics
with respect to predicting values in new unseen environments

2. causal inference = intervention analysis

in genomics (for yeast or plants):
if we would make an intervention at a single (or many) gene(s),
what would be its (their) effect on a response of interest?

want to infer/predict such effects without actually doing the intervention
e.g. from observational data (cf. Pearl; Spirtes, Scheines & Glymour)
(from observations of a "steady-state system")
or from observational and interventional (heterogeneous) data
$\rightsquigarrow$ want to predict unseen interventions

we need a model, of course! (one which is good/"justifiable")

2. causal inference = intervention analysis

in genomics (for yeast or plants):
if we would make an intervention at a single (or many) gene(s),
what would be its (their) effect on a response of interest?

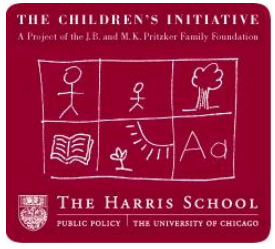
want to infer/predict such effects without actually doing the intervention
e.g. from observational data (cf. Pearl; Spirtes, Scheines & Glymour)
(from observations of a "steady-state system")
or from observational and interventional (heterogeneous) data
$\rightsquigarrow$ want to predict unseen interventions

we need a model, of course! (one which is good/"justifiable")

Example: Policy making

James Heckman: Nobel Prize Economics 2000

e.g.:
“Pritzker Consortium on Early Childhood Development identifies when and how child intervention programs can be most influential”
$\rightsquigarrow$ predict what happens if child would be assigned to an educational program “X” for which we have no data



Example: Flowering of Arabidopsis Thaliana

phenotype/response variable of interest:
$Y =$ days to bolting (flowering)
"covariates" $X =$ gene expressions from $p = 21'326$ genes

goal: based on observational/interventional data,
predict the effect of knocking-out a new single gene on the response variable Y

and we can validate the prediction by doing randomized follow-up experiments afterwards
(Stekhoven, Moraes, Sveinbjörnsson, Hennig, Maathuis & PB, 2012)

in both

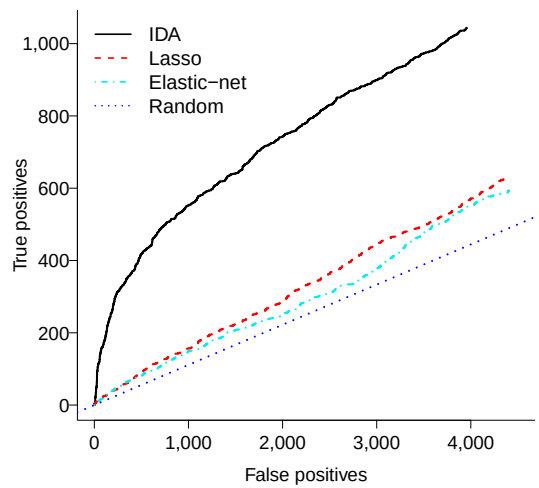
- prediction in heterogeneous environments
- causal inference

$\rightsquigarrow$ prediction for new unseen scenarios/environments

$\rightsquigarrow$ “equivalence” of problems!

~~REGRESSION~~

validated with follow-up biological experiments



~~REGRESSION~~

because: for

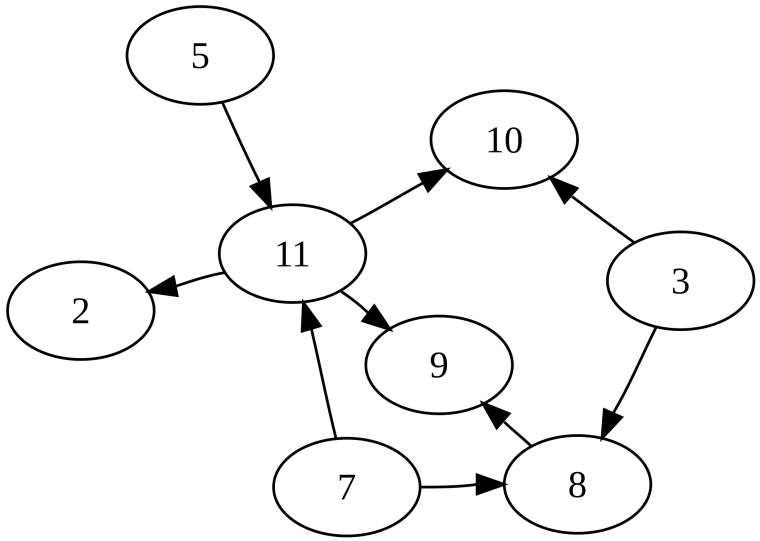
$$Y = \sum_{j=1}^{p} \beta_j X^{(j)} + \varepsilon$$

β_j measures effect of $X^{(j)}$ on Y when
keeping all other variables $\{X^{(k)};\ k \neq j\}$ fixed

but when doing an intervention at a gene $\rightsquigarrow$ some/many other genes might change as well and cannot be kept fixed

Causality, Graphical and Structural equation models

late 1980s: Pearl; Spirtes, Glymour, Scheines; Dawid; Lauritzen;. . .



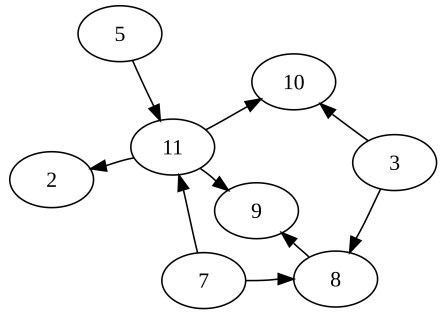
"definition" of causality:

- direct causal variables for Y: the parental variables of Y
- total causal effect of $X^{(j)}$ on Y:
  intervention or "treatment" effect of $X^{(j)}$ on Y
  $\mathrm{do}(X^{(j)} = x)$: the effect on Y when setting $X^{(j)} = x$

  $\rightsquigarrow$ sum up directed paths ("edge weights") from $X^{(j)}$ to Y

variables $X_1, \ldots, X_{p+1}$ ($X_{p+1} = Y$ is the response of interest)

directed acyclic graph (DAG) D^0 encoding the true underlying causal influence diagram



structural equation model (SEM):

$$X_j \leftarrow f_j^0(X_{\mathrm{pa}_{D^0}(j)}, \varepsilon_j), \quad j = 1, \ldots, p+1,$$
$$\varepsilon_1, \ldots, \varepsilon_{p+1} \text{ independent}$$

e.g. linear $$X_j \leftarrow \sum_{k \in \mathrm{pa}_{D^0}(j)} \beta_{jk}^0 X_k + \varepsilon_j, \quad j = 1, \ldots, p+1$$

causal variables for $Y = X_{p+1}$: $S^0 = \{k;\ k \in \mathrm{pa}_{D^0}(Y)\}$

severe issues of identifiability !

given distribution(s) generating the data: typically cannot identify the true DAG D^0 and the parental set S^0
examples:

$$(X, Y) \sim \mathcal{N}_2(0, \Sigma)$$

X causes Y

Y causes X

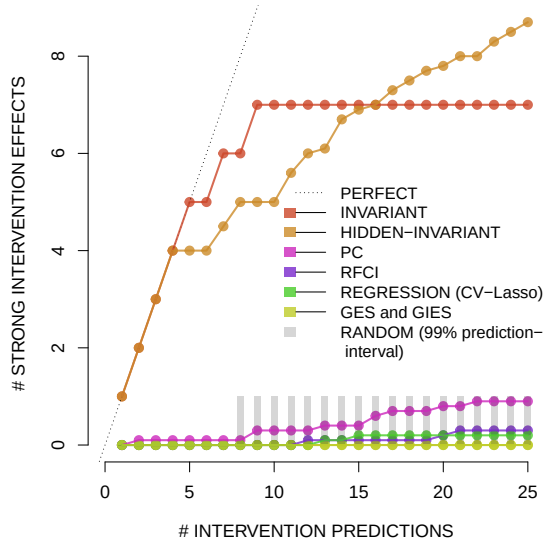
An equivalence class can be uniquely represented by a completed partially directed acyclic graph (CPDAG)

agenda for estimation: based on observ. or observ./interv. data
(Chickering, 2002; Shimizu, 2005; Kalisch & PB, 2007;...)

1. estimate the Markov equivalence class of DAGs
   severe issues of identifiability !
2. derive causal variables: the ones which are causal in all DAGs from; derive bounds for causal effects (Maathuis, Kalisch & PB, 2009)

drawbacks:

- rather unstable and “doesn’t really work”



I : invariant prediction method

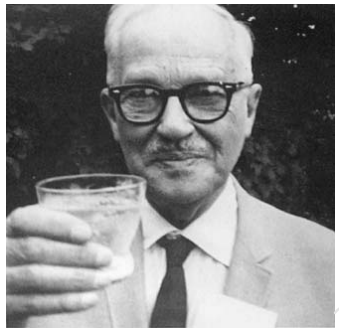
H: invariant prediction with some hidden variables

- no confidence statements
- is tailored for very specidifc interventions (experimental conditions) only

goals:

1. construction of confidence statements for causal var. S^0
   (without knowing the structure of the underlying graph)
2. deal with "unspecified" heterogeneous/interv. data
   general

NOT or AVOIDING
– graphical model fitting
– potential outcome models



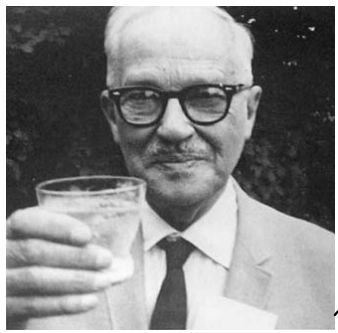
↝ Neyman's master thesis 1923!

goals:

1. construction of confidence statements for causal var. S^0
   (without knowing the structure of the underlying graph)

2. deal with $\underbrace{\text{“unspecified”}}_{\text{general}}$ heterogeneous/interv. data

NOT or AVOIDING
– graphical model fitting
– potential outcome models



$\rightsquigarrow$ Neyman's master thesis 1923!

Causal inference using invariant prediction

Peters, PB and Meinshausen (2016)

a main message:

causal structure/components remain the same for different sub-populations

while the non-causal components can change across sub-populations

thus:

$\rightsquigarrow$ look for "stability" of structures among different sub-populations

Causal inference using invariant prediction

Peters, PB and Meinshausen (2016)

a main message:

causal structure/components remain the same for different sub-populations

while the non-causal components can change across sub-populations

thus:

$\rightsquigarrow$ look for "stability" of structures among different sub-populations

Heterogeneous data

$$(X^e, Y^e) \sim F^e, \; e \in \underbrace{\mathcal{E}}_{\text{space of observed experimental conditions}}$$

example 1: $\mathcal{E} = \{1, 2\}$ encoding observational (1) and all potentially unspecific interventional data (2)

example 2: $\mathcal{E} = \{1, 2\}$ encoding observational data (1) and (repeated) data from one specific intervention (2)

example 3: $\mathcal{E} = \{1, 2, 3\}$ ... or $\mathcal{E} = \{1, 2, 3, \ldots, 26\}$ ...

do not need data from carefully designed (randomized) experiments

Invariance Assumption (w.r.t. $\mathcal{E}$)
there exists $S^* \subseteq \{1, \ldots, p\}$ such that:

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is invariant across } e \in \mathcal{E}$$

for linear model setting:
there exists a vector γ^* with $\text{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\forall e \in \mathcal{E}: \quad Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*}$$

$\varepsilon^e \sim F_\varepsilon$ the same for all e

X^e has an arbitrary distribution, different across e

$\gamma^*,\ S^*$ is interesting in its own right!

namely the parameter and structure which remain invariant
across experimental settings, or across heterogeneous groups

Invariance Assumption (w.r.t. $\mathcal{E}$)
there exists $S^* \subseteq \{1, \ldots, p\}$ such that:

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is invariant across } e \in \mathcal{E}$$

for linear model setting:
there exists a vector γ^* with $\text{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\forall e \in \mathcal{E}: \quad Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*}$$

$\varepsilon^e \sim F_\varepsilon$ the same for all e

X^e has an arbitrary distribution, different across e

$\gamma^*,\ S^*$ is interesting in its own right!

namely the parameter and structure which remain invariant across experimental settings, or across heterogeneous groups

Invariance Assumption w.r.t. $\mathcal{F}$

$$\text{where } \mathcal{F} \underbrace{\supset}_{\text{much larger}} \mathcal{E}$$

now: the set S^* and corresponding regression parameter γ^* are for a much larger class of environments than what we observe!

$\rightsquigarrow$

$$\gamma^*,\ S^* \text{ is even more interesting in its own right!}$$

since it says something about unseen new environments!

Link to causality

Invariance Assumption w.r.t. any space of environments $\mathcal{G}$:

there exists S^* such that $\mathcal{L}(Y^e|X^e_{S^*})$ is invariant across $e \in \mathcal{G}$

Proposition (Peters, PB & Meinshausen, 2016)
Assume structrual equation model (SEM)

$$X_1 \leftarrow f_1^0(X_{\mathrm{pa}(1)}, \varepsilon_1),$$
$$X_2 \leftarrow f_2^0(X_{\mathrm{pa}(2)}, \varepsilon_2),$$
$$\ldots$$
$$Y \leftarrow f_Y^0(X_{\mathrm{pa}(Y)}, \varepsilon_Y)$$

Assume that $\mathcal{G}$ does not affect the structural equation for Y:

e.g. linear SEM: $Y^e \leftarrow \sum_{k \in \mathrm{pa}(Y)} \underbrace{\beta_{Yk}}_{\forall e} X_k^e + \underbrace{\varepsilon_Y^e}_{\sim F_\varepsilon \forall e \in \mathcal{G}}$

Then: $\underbrace{S^0 = \mathrm{pa}(Y)}_{\text{causal var.}}$ satisfies the Invariance Assumption w.r.t. $\mathcal{G}$

can take $\mathcal{G} = \mathcal{E}$, $\mathcal{G} = \mathcal{F} =$ all environments, ...

Link to causality

Invariance Assumption w.r.t. any space of environments $\mathcal{G}$:

there exists S^* such that $\mathcal{L}(Y^e | X^e_{S^*})$ is invariant across $e \in \mathcal{G}$

Proposition (Peters, PB & Meinshausen, 2016)
Assume structrual equation model (SEM)

$$
\begin{aligned}
X_1 &\leftarrow f_1^0(X_{\mathrm{pa}(1)}, \varepsilon_1), \\
X_2 &\leftarrow f_2^0(X_{\mathrm{pa}(2)}, \varepsilon_2), \\
&\dots \\
Y &\leftarrow f_Y^0(X_{\mathrm{pa}(Y)}, \varepsilon_Y)
\end{aligned}
$$

Assume that $\mathcal{G}$ does not affect the structural equation for Y:

e.g. linear SEM: $Y^e \leftarrow \sum_{k \in \mathrm{pa}(Y)} \underbrace{\beta_{Yk}}_{\forall e} X_k^e + \underbrace{\varepsilon_Y^e}_{\sim F_\varepsilon \forall e \in \mathcal{G}}$

Then: $\underbrace{S^0 = \mathrm{pa}(Y)}_{\text{causal var.}}$ satisfies the Invariance Assumption w.r.t. $\mathcal{G}$

can take $\mathcal{G} = \mathcal{E}$, $\mathcal{G} = \mathcal{F} =$ all environments, ...

the causal variables lead to invariance (of conditional distr.)

w.r.t. “all” possible environments

the Proposition has been known for a long time in causality
(Haavelmo, 1944; Aldrich, 1989; Hoover, 1990; ... Dawid and Didelez, 2010)

causal structure (parental variables) $\Longrightarrow$ invariance

the new thing (surprisingly!) will be the reverse relation:

causal structure (parental variables) $\Longleftarrow$ invariance

the causal variables lead to invariance (of conditional distr.)
w.r.t. "all" possible environments

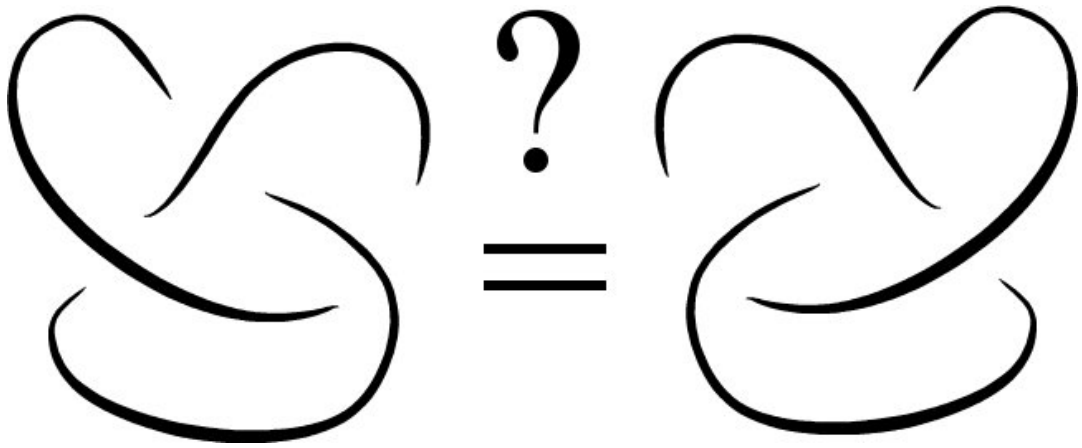
the Proposition has been known for a long time in causality
(Haavelmo, 1944; Aldrich, 1989; Hoover, 1990; ... Dawid and Didelez, 2010)

$$\text{causal structure (parental variables)} \implies \text{invariance}$$

the new thing (surprisingly!) will be the reverse relation:

$$\text{causal structure (parental variables)} \impliedby \text{invariance}$$

invariance – an important mathematical and scientific concept

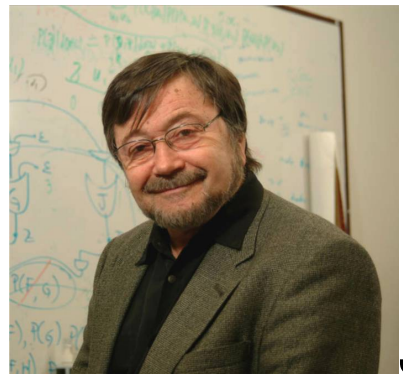


recap: main assumptions implying that
the causal variables lead to invariance

- a structural equation model
- $\mathcal{E}$ (or $\mathcal{F} \supset \mathcal{E}$) does not affect structural equation for Y

this assumption holds for example for:

- do-intervention (Pearl) at variables different than Y

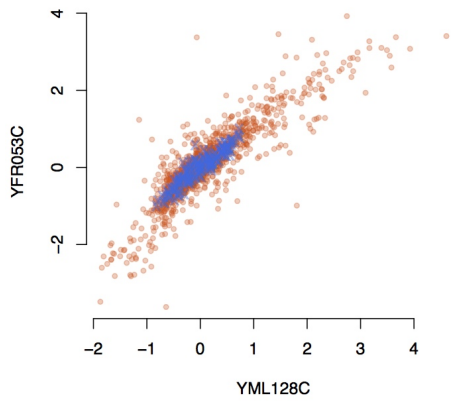

Judea Pearl

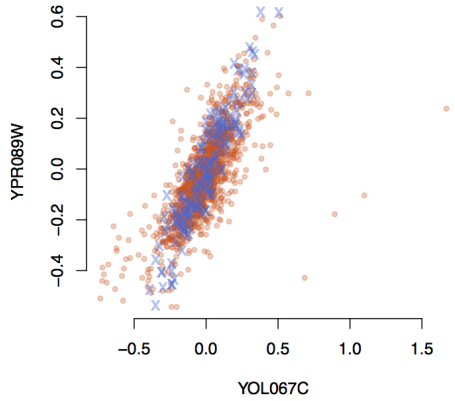
- noise (or "soft") intervention (Eberhardt & Scheines, 2007) at variables different than Y

Invariance Assumption : plausible to hold with real data

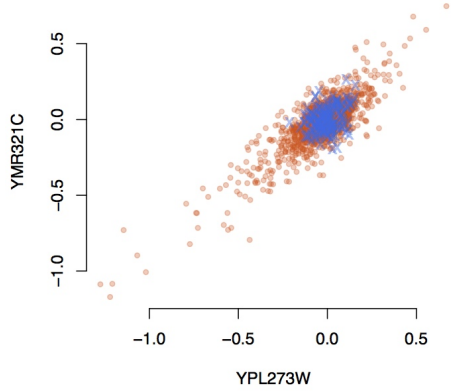
two-dimensional conditional distributions of observational (blue) and interventional (orange) data
(no intervention at displayed variables X, Y)

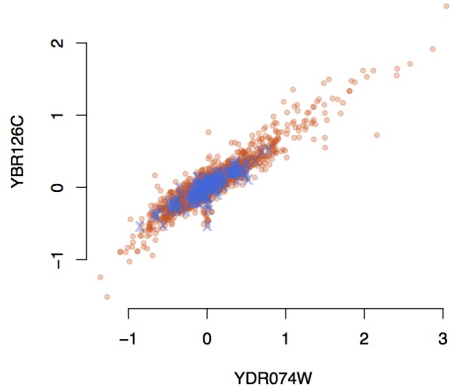
seemingly
no invariance
of conditional d.





plausible
invariance
of conditional d.





A procedure for inferring S^0: population case

require and exploit the Invariance Assumption (w.r.t. $\mathcal{E}$)

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ the same across } e \in \mathcal{E}$$

for linear model: consider hypothesis

$H_{0,S}(\mathcal{E})$: there exists γ with $\text{supp}(\gamma) = S$ and
there exists F_ε such that $\forall\ e \in \mathcal{E}$:
$Y^e = X^e\gamma + \varepsilon^e,\ \varepsilon^e \perp X^e_S,\ \varepsilon^e \sim F_\varepsilon$ the same for all e

i.e. $H_{0,S}(\mathcal{E})$ holds $\leftrightarrow$ Invariance Assumption holds for set S

and there might be many such S...

identifiable causal variables/predictors under $\mathcal{E}$:

is defined as the set $S(\mathcal{E})$, where

$$S(\mathcal{E}) = \bigcap \{ \underbrace{S;\ H_{0,S}(\mathcal{E}) \text{ holds}}_{\text{Invariance Assumption holds for } S} \}$$

the intersection of all sets S where Inv. Ass. holds

for any S^* satisfying the Invariance Assumption we have:

$$S(\mathcal{E}) \subseteq S^*$$

and this is key to obtain confidence statements for identifiable causal variables

identifiable causal variables/predictors under $\mathcal{E}$:

is defined as the set $S(\mathcal{E})$, where

$$S(\mathcal{E}) = \bigcap \{ \underbrace{S;\ H_{0,S}(\mathcal{E}) \text{ holds}}_{\text{Invariance Assumption holds for } S} \}$$

the intersection of all sets S where Inv. Ass. holds

for any S^* satisfying the Invariance Assumption we have:

$$S(\mathcal{E}) \subseteq S^*$$

and this is key to obtain confidence statements for identifiable causal variables

we have by definition:

$$S(\mathcal{E}) \nearrow \text{ as } \mathcal{E} \nearrow$$

with

- more interventions
- more “heterogeneity”
- more “diversity in complex data”

  we can identify more causal variables

question: when is $S(\mathcal{E}) = S^0$?

answer not of primary importance
(see later)

Theorem (Peters, PB and Meinshausen, 2016)

$$S(\mathcal{E}) = S^0 \quad = \text{(parental set of } Y \text{ in the causal DAG)}$$

if there is:

- a single do-intervention for each variable other than Y and $|\mathcal{E}| = p$
- a single noise intervention for each variable other than Y and $|\mathcal{E}| = p$
- a simultaneous noise intervention and $|\mathcal{E}| = 2$

the conditions can be relaxed such that it is not necessary to intervene at all the variables

Statistical confidence sets for causal predictors

"the finite sample version of $S(\mathcal{E}) = \bigcap_S \{S;\ H_{0,S}(\mathcal{E}) \text{ is true}\}$"

for "any" $S \subseteq \{1, \ldots, p\}$:
test whether $H_{0,S}(\mathcal{E})$ is accepted or rejected

$$\hat{S}(\mathcal{E}) = \bigcap_S \{H_{0,S} \text{ accepted at level } \alpha\}$$

for $H_{0,S}(\mathcal{E})$:
test constancy of regression param. and of residual error distr. across $e \in \mathcal{E}$

weaken this $\tilde{H}_{0,S}(\mathcal{E})$:
test constancy of regression param. and of standard deviation of residual error across $e \in \mathcal{E}$

known since a long time how to do this:
assume Gaussian errors
$\rightsquigarrow$ an exact test with an F-distribution under $\tilde{H}_{0,S}(\mathcal{E})$

$$\hat{S}(\mathcal{E}) = \bigcap_S \{\tilde{H}_{0,S} \text{ accepted at level } \alpha\}$$

for some significance level $0 < \alpha < 1$

no multiple testing adjustment is needed!

method is called: ICP = Invariant Causal Prediction

going through all sets S?

$$\hat{S}(\mathcal{E}) = \bigcap_S \{\tilde{H}_{0,S} \text{ accepted at level } \alpha\}$$

for some significance level $0 < \alpha < 1$

no multiple testing adjustment is needed!

method is called: ICP = Invariant Causal Prediction

going through all sets S?

going through all sets S? in the worst case: yes

1. start with $S = \emptyset$: if $H_{0,\emptyset}(\mathcal{E})$ accepted $\implies \hat{S}(\mathcal{E}) = \emptyset$
2. consider small sets S of cardinality $1, 2, \ldots$
   and construct corresponding intersections $S_\cap$ with
   previously considered accepted sets S ($H_{0,S}(\mathcal{E})$ accepted)

   $$\text{for } S \text{ with } H_{0,S} \text{ accepted :}$$
   $$S_\cap \leftarrow S_\cap \cap S$$

   if intersection $S_\cap = \emptyset \implies \hat{S}(\mathcal{E}) = \emptyset$
   if not:

   $$\text{discard all } S \text{ with } S \supseteq S_\cap$$

   and continue with the remaining sets
3. for large p:
   restrict search space by variables from Lasso regression;
   need a faithfulness assumption (and sparsity and assumptions on X^e for justification)

Theorem (Peters, PB and Meinshausen, 2016)

assume: linear model, Gaussian errors
$\mathcal{E}$ does not affect structural equation for Y

Then:

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq S^0] \geq 1 - \alpha : \text{ confidence w.r.t. true causal var.}$$

"on the safe side" (conservative)

we do not need to care about identifiability: if the effect is not identifiable, the method will not wrongly claim an effect

Theorem (Peters, PB and Meinshausen, 2016)

assume: linear model, Gaussian errors
$\mathcal{E}$ does not affect structural equation for Y

Then:

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq S^0] \geq 1 - \alpha : \text{ confidence w.r.t. true causal var.}$$

"on the safe side" (conservative)

we do not need to care about identifiability: if the effect is not identifiable, the method will not wrongly claim an effect

“the first” result on
frequentist statistical confidence for potentially non-identifiable causal predictors when structure is unknown

(route via graphical modeling for confidence sets seems awkward)

leading to (hopefully) more
reliable causal inferential statements

how do we know whether
$\mathcal{E}$ is not affecting structural equation for Y?

if $\mathcal{E}$ does affect structural equation for Y

$\rightsquigarrow$

"robustness" of our procedure

– no causal statements
– no false positives
– conservative, but on the safe side

how do we know whether
$\mathcal{E}$ is not affecting structural equation for Y?

if $\mathcal{E}$ does affect structural equation for Y

$\rightsquigarrow$

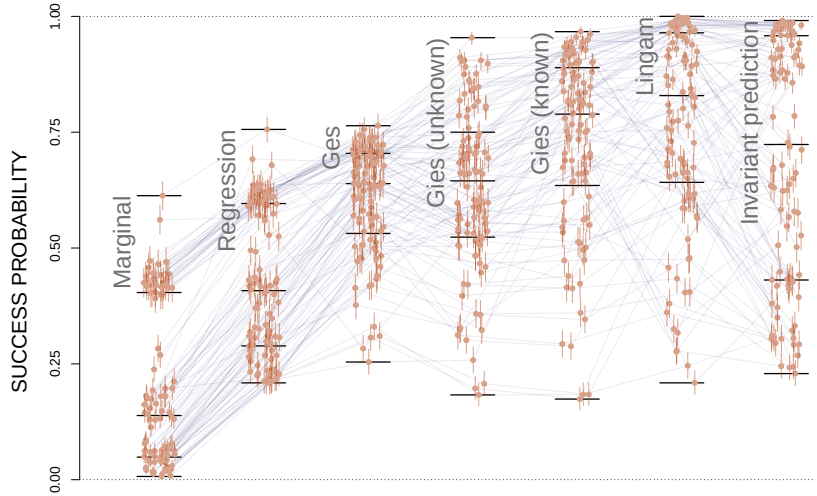
"robustness" of our procedure

– no causal statements
– no false positives
– conservative, but on the safe side

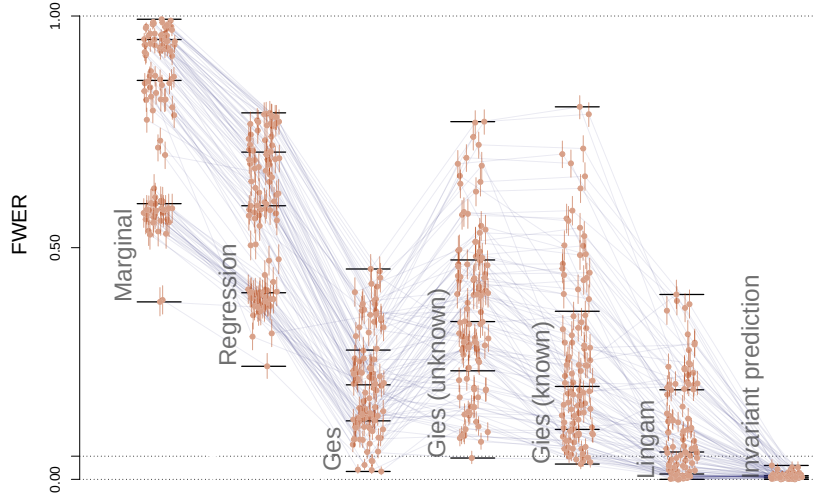
Empirical results: simulations

100 different scenarios, 1000 data sets per scenario:

$$|\mathcal{E}| = 2,\ n_{obs} = n_{interv} \in \{100, \ldots, 500\},\ p \in \{5, \ldots, 40\}$$

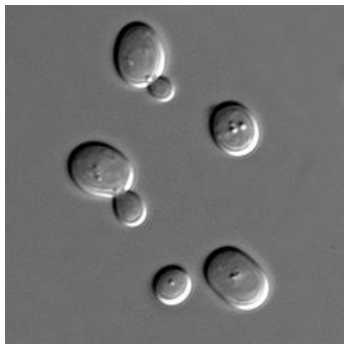


power to
detect causal predictors



familywise error rate:
$\mathbb{P}[\hat{S}(\mathcal{E}) \not\subseteq S^0]$, aimed at 0.05

Single gene deletion experiments in yeast



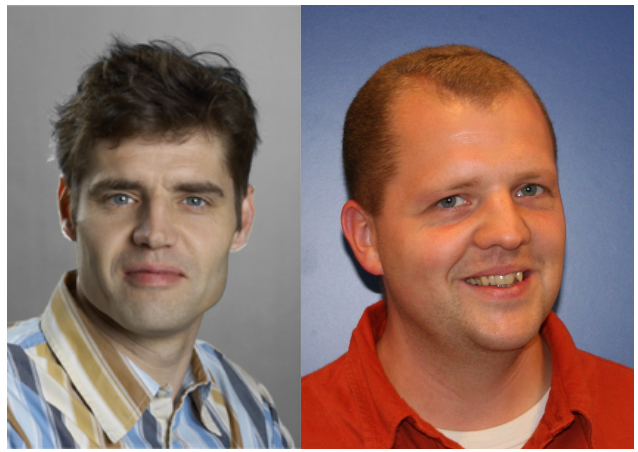
$p = 6170$ genes
response of interest: $Y =$ expression of first gene
"covariates" $X =$ gene expressions from all other genes

and then
response of interest: $Y =$ expression of second gene
"covariates" $X =$ gene expressions from all other genes

and so on

infer/predict the effects of unseen/new single gene deletions on all other genes
that is: make predictions for
new observations from new probability distributions

collaborators:
Frank Holstege, Patrick Kemmeren et al. (Utrecht)



data from modern technology

Kemmeren, ..., and Holstege (Cell, 2014)

Kemmeren et al. (2014):
genome-wide mRNA expressions in yeast: $p = 6170$ genes

- $n_{obs} = 160$ "observational" samples of wild-types
- $n_{int} = 1479$ "interventional" samples
  each of them corresponds to a single gene deletion strain

for our method: we use $|\mathcal{E}| = 2$ (observational and interventional data)

training-test data splitting:

- training set: all observational and 2/3 of interventional data
- test set: other 1/3 of gene deletion interventions
  predicted effects of these interventions are validated
- repeat this for the three blocks of interventional test data

multiplicity adjustment:
since ICP is used 6170 times (once for every response var.) we use coverage $1 - \alpha/6170$ with $\alpha = 0.05$

Results for inferring causal variables

8 genes are significant ($\alpha = 0.05$ level) causal variables
(each of the 8 genes "causes" one other gene)

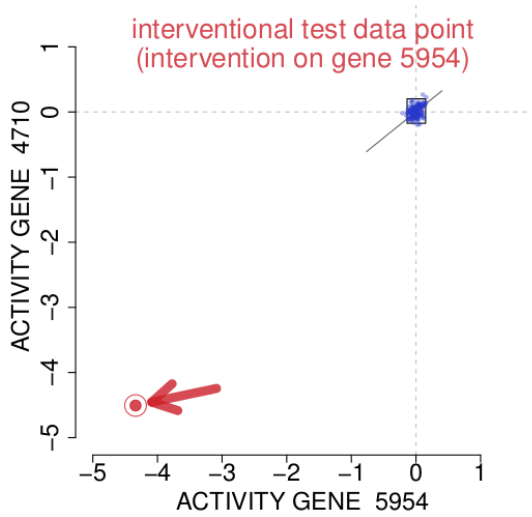
not many findings...
but we use a stringent criterion with Bonferroni corrected
$\alpha/6170 = 0.05/6170$ to control the familywise error rate

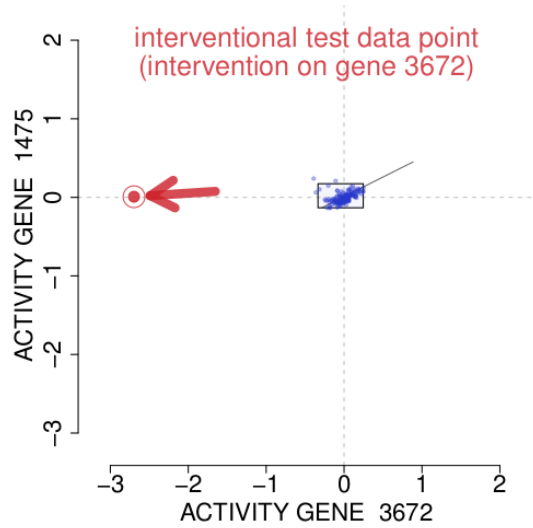
and ICP might be conservative (as discussed before)

8 genes are significant ($\alpha = 0.05$ level) causal variables

validation:
thanks to the intervention experiments (in the test data) we can validate the method(s)

SIE = the observed response value associated to an intervention is in the 1%- or 99% tail of the observational data





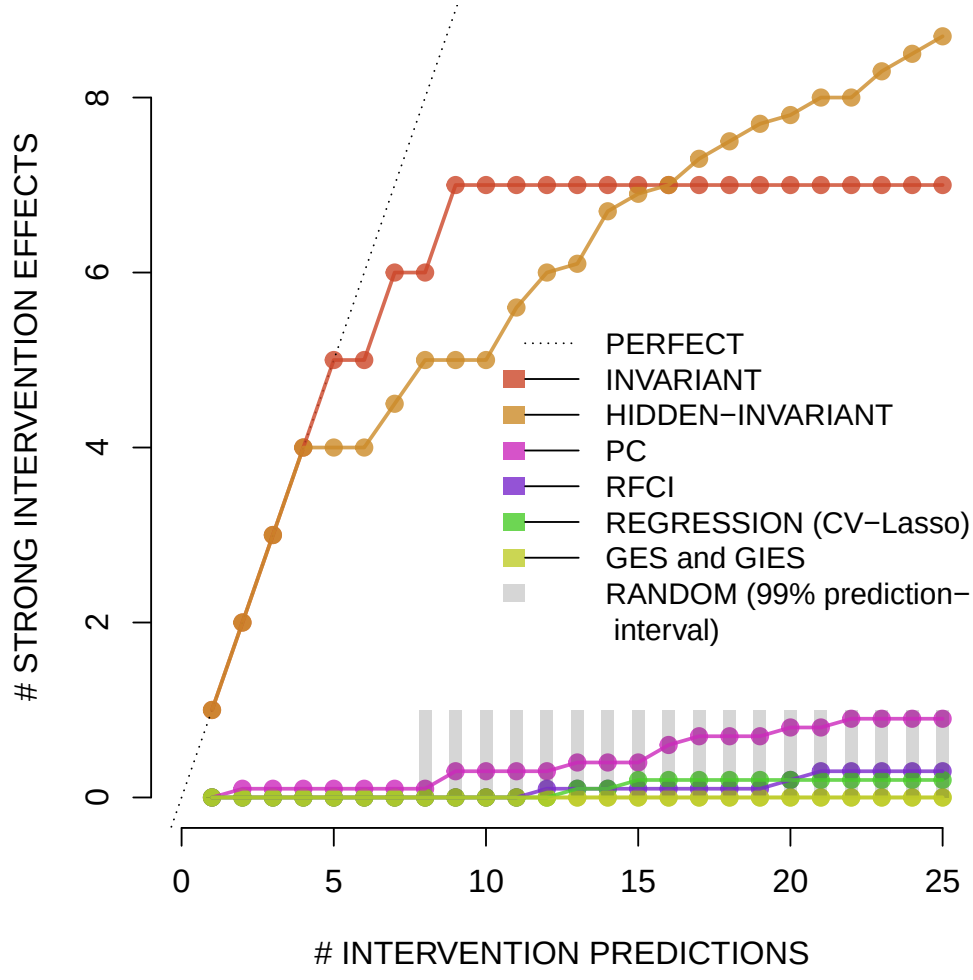
$\rightsquigarrow$ a very stringent conservative definition of a true positive intervention effect

8 genes are significant ($\alpha = 0.05$ level) causal variables

method	invar.pred.	GIES	PC-IDA	marg.corr.	rand.guess.
no. true pos. (out of 8)	6	2	2	2	*

*: quantiles for selecting true positives among 7 random draws
2 (95%), 3 (99%)

$\rightsquigarrow$ our invariant prediction method has most power !
and it should exhibit control against false positive selections



I : invariant prediction method

H: invariant prediction with some hidden variables

Validation (Meinshausen, Hauser, Mooij, Peters, Versteeg & PB, 2016) with intervention experiments: strong intervention effect (SIE) with `yeastgenome.org` database: scores A-F

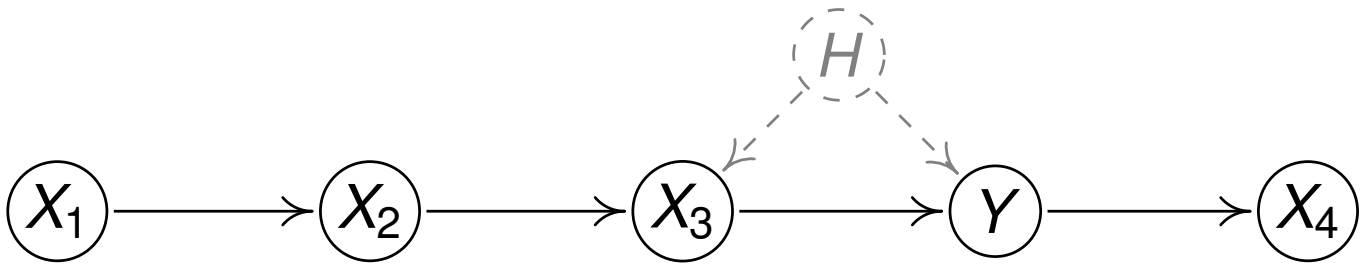
rank	cause	effect	SIE	A	B	C	D	E	F
1	YMR104C	YMR103C	✓						
2	YPL273W	YMR321C	✓						
3	YCL040W	YCL042W	✓						
4	YLL019C	YLL020C	✓						
5	YMR186W	YPL240C	✓	✓	✓	✓	✓		✓
6	YDR074W	YBR126C		✓	✓	✓	✓	✓	✓
7	YMR173W	YMR173W-A	✓						
8	YGR162W	YGR264C							
9	YOR027W	YJL077C	✓						
10	YJL115W	YLR170C							
11	YOR153W	YDR011W		✓	✓				
12	YLR270W	YLR345W							
13	YOR153W	YBL005W							
14	YJL141C	YNR007C							
15	YAL059W	YPL211W							
16	YLR263W	YKL098W							
17	YGR271C-A	YDR339C							
18	YLL019C	YGR130C							
19	YCL040W	YML100W							
20	YMR310C	YOR224C							

SIE: correctly predicting a strong intervention effect which is in the 1%- or 99% tail of the observational data

Robustness

remember:

- if model is not correct exhibiting e.g. nonlinearities
  $\rightsquigarrow$ loss of power, but controlling false positives is still OK
- if Invariance Assumption does not hold
  $\rightsquigarrow$ loss of power, but controlling false positives is still OK
- hidden variables
  $\rightsquigarrow$ the method might pick up ancestors of Y

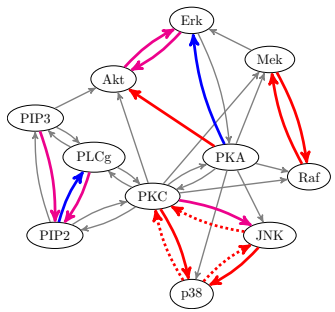


e.g. X_2 which still exhibits a total intervention/causal effect (and hence is interesting for the gene perturbation experiments)

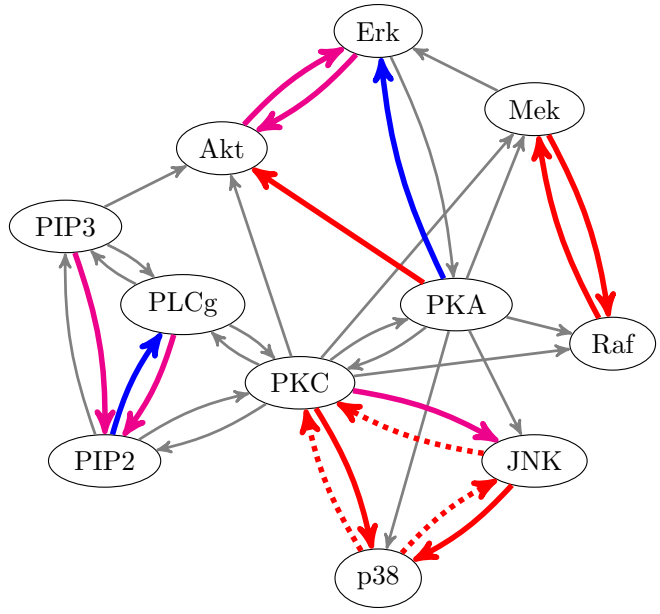
Flow cytometry data (Sachs et al., 2005)

- $p = 11$ abundances of chemical reagents
- 8 different environments (not "well-defined" interventions)
  (one of them observational; 7 different reagents added)
- each environment contains $n_e \approx 700 - 1'000$ samples

goal:
recover network of causal relations (linear SEM)



approach: "pairwise" invariant causal prediction
(one variable the response Y; the other 10 the covariates X;
do this 11 times with every variable once the response)



blue edges: only invariant causal prediction approach (ICP)
red: only ICP allowing hidden variables and feedback
purple: both ICP with and without hidden variables
solid: all relations that have been reported in literature
broken: new findings not reported in the literature

$\rightsquigarrow$ reasonable consensus with existing results
but no real ground-truth available
serves as an illustration that we can work with "vaguely defined interventions"

Concluding thoughts

generalize Invariance Assumption and statistical testing to nonparametric/nonlinear models
in particular also additive models

$$\forall e \in \mathcal{E}: \; Y^e = f^*(X^e_{S^*}) + \varepsilon^e, \; \varepsilon^e \sim F_\varepsilon, \; \varepsilon^e \perp X_{S^*}$$
$$\forall e \in \mathcal{E}: \; Y^e = \sum_{j \in S^*} f^*_j(X^e_j) + \varepsilon^e, \; \varepsilon^e \sim F_\varepsilon, \; \varepsilon^e \perp X_{S^*}$$

the statistical significance testing becomes more difficult
improved identifiability with nonlinear SEMs (Mooij et al., 2009)

provocative next step:
how about using “Big Data” when $\mathcal{E}$ is unknown?

that is: learn $\mathcal{E}$ from data

$\rightsquigarrow$ partition $\mathcal{E}$ to maximize the number of confident detections
(wrong partitions will not destroy type I error control)

- need to adjust for searching for best partition
- much easier for (time-ordered) data

  $\rightsquigarrow$ some kind of change point/segmentation problem

(work in progress by Pfister & PB)

further issues:

- feedback loops in causal influence diagram
  (Rothenhäusler, Heinze, Peters & Meinshausen, 2015)
- hidden variables
  (Rothenhäusler, Heinze, Peters & Meinshausen, 2015)
- dynamic processes (with applications in economics, finance, neuroscience,...)
- ...

causal components remain the same for
different sub-populations or experimental settings

$\rightsquigarrow$ useful for

- causal inference with confidence statements
  (as illustrated in this talk)
- prediction in heterogeneous environments (in progress)

$\rightsquigarrow$ exploit the power of heterogeneity in complex data!

Thank you!

Software

R-package: `pcalg`
(Kalisch, Mächler, Colombo, Maathuis & PB, 2010–2015)

R-package: `InvariantCausalPrediction` (Meinshausen, 2014)

References to some of our own work:

- Peters, J., Bühlmann, P. and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals (with discussion). J. Royal Statistical Society: Series B 78, 947-1012.
- Meinshausen, N., Hauser, A. Mooij, J., Peters, J., Versteeg, P. and Bühlmann, P. (2016). Causal inference from gene perturbation experiments: methods, software and validation. Proc. Nat. Acad. Sci. USA 113, 7361-7368.
- Hauser, A. and Bühlmann, P. (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society: Series B, 77, 291-318.
- Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research 13, 2409-2464.
- Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H. and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47 (11), 1-26.
- Stekhoven, D.J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M.H. and Bühlmann, P. (2012). Causal stability ranking. Bioinformatics 28, 2819-2823.
- Maathuis, M.H., Colombo, D., Kalisch, M. and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. Nature Methods 7, 247-248.
- Maathuis, M.H., Kalisch, M. and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. Annals of Statistics 37, 3133-3164.