Invariance, Causality and novel Robustness

Peter Bühlmann

based on collaborations with

Jonas Peters
Univ. Copenhagen

Nicolai Meinshausen
ETH Zürich

Dominik Rothenhäusler
now UC Berkeley

Causality – Robustness

we have been working on the former “exotic” problem for a while

but it turns out that there are connections to the latter

a nutshell view of Robust Optimization

distributionally robust optimization
(Ben-Tal, El Ghaoui & Nemirovski, 2009; ...
e.g. Sinha, Namkoong & Duchi, 2017)

$$\text{argmin}_\beta \max_{P \in \mathcal{P}} \underbrace{\mathbb{E}_P[\ell(X, Y; \beta)]}_{\text{e.g. } \mathbb{E}_P|Y - X^T\beta|^2}$$

good performance under adversarial distributions

typically

$$\mathcal{P} = \{P; \quad \underbrace{d(P, P_0)}_{\text{e.g. Wasserstein distance}} \quad \leq \rho\}$$

often $\quad P_0 = \hat{P}$ ($=$ empirical dist.)

a nutshell view of Robust Optimization

distributionally robust optimization
(Ben-Tal, El Ghaoui & Nemirovski, 2009; ...
e.g. Sinha, Namkoong & Duchi, 2017)

$$\text{argmin}_\beta \max_{P \in \mathcal{P}} \underbrace{\mathbb{E}_P[\ell(X, Y; \beta)]}_{\text{e.g. } \mathbb{E}_P|Y - X^T\beta|^2}$$

good performance under adversarial test sample distributions

typically

$$\mathcal{P} = \{P; \quad \underbrace{d(P, P_0)}_{\text{e.g. Wasserstein distance}} \quad \le \rho\}$$

often $\quad P_0 = \hat{P}$ ($=$ empirical dist.)

Robust Statistics

Huber's (1964) celebrated minimax result (for location)

"... there is a saddle point... "

$$\min_{\hat{\beta}\in\mathcal{M}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[(Y_{\mathrm{out}} - \hat{\beta})^2]$$

is achieved by the Huber estimator

$\rightsquigarrow$ it is an estimator with another loss function ρ_{Huber} replacing the worst case L_2-loss

good performance under contaminated distributions

$\mathcal{P}$ is a neighborhood of a Gaussian reference distribution P_0

Huber's (1964) celebrated minimax result (for location)

"... there is a saddle point... "

$$\min_{\hat{\beta}\in\mathcal{M}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[(Y_{\text{out}} - \hat{\beta})^2]$$

is achieved by the Huber estimator
$\leadsto$ it is an estimator with another loss function ρ_{Huber}
replacing the worst case L_2-loss
good performance under contaminated training sample distr.

$\mathcal{P}$ is a neighborhood of a Gaussian reference distribution P_0

A “general” robustness view

achieve stability (or “near invariance”, see later)
for a class of “meaningful/interesting” distributions $\mathcal{P}$

$\mathcal{P}$ is not necessarily a
– “neighborhood”
– “ball of a certain radius”

try to capture the “interesting directions” for which we want to achieve stability (perhaps specific to different data analysis problems)

as in classical statistical robustness, robust optimization, adversarial training,...

but also Causality can be looked at from this viewpoint!

of course, such “general robustness” is not new:
Tukey (1960), Hodges& Lehmann (1963), Huber (1964), Bickel (1964), Hampel (1968), ... , Soyster (1973), ..., Haavelmo (1943), ...

Causality

the word “causal” is very ambitious...

perhaps too ambitious...
but we aim at least at doing something “ more suitable” than standard regression or classification

The gold standard

a randomized control trial (RCT)

What can we say without RCTs? $\longrightarrow$ Prediction!

causality is about giving a quantitative answer to a

- "what if I do question"
- a "what if I perturb question"

but without having data on such a question

Predicting

the outcome of an unobserved manipulation

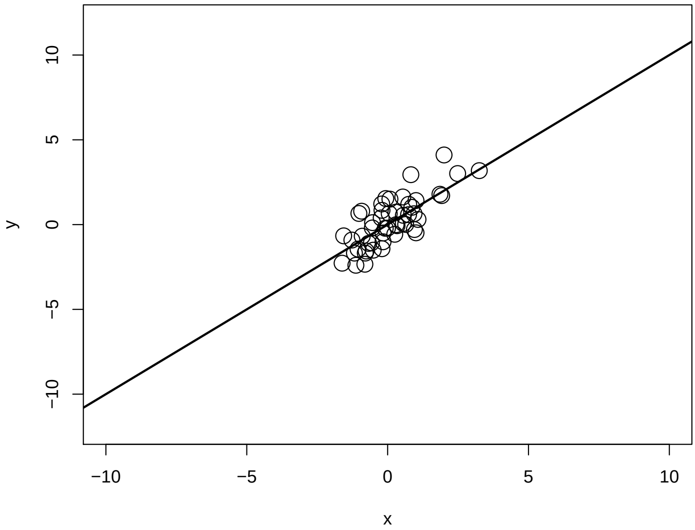
(and it is also about predictive robustness) many modern applications are faced with such prediction tasks:

- genomics: what would be the effect of knocking down (the activity of) a gene on the growth rate of a plant?

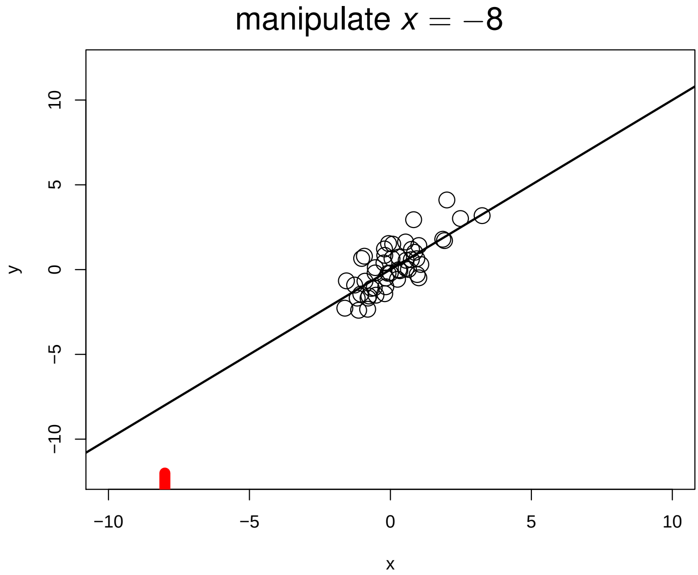
  

  we want to predict this without any data on such a gene knock-out (e.g. no data for this particular perturbation)
- E-commerce: what would be the effect of showing person "*XYZ*" an advertisement on social media?
  no data on such an advertisement campaign for "*XYZ*" or persons being similar to "*XYZ*"
- etc.

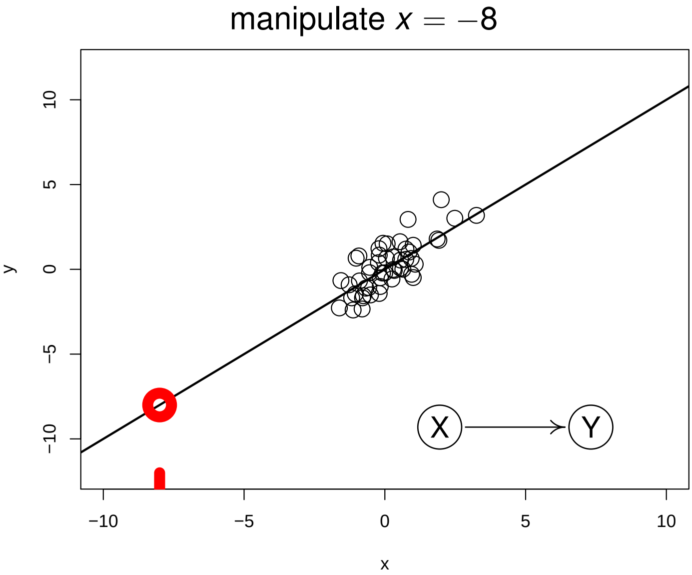
Predicting a potential outcome (synthetic data)



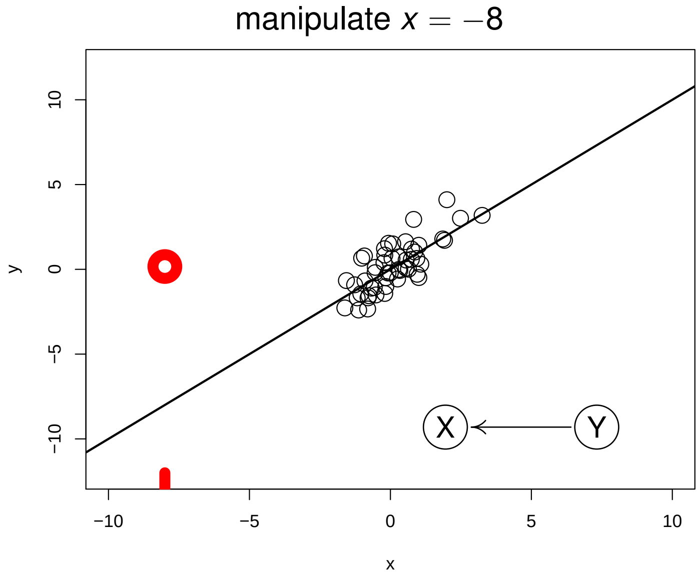
Predicting a potential outcome (synthetic data)



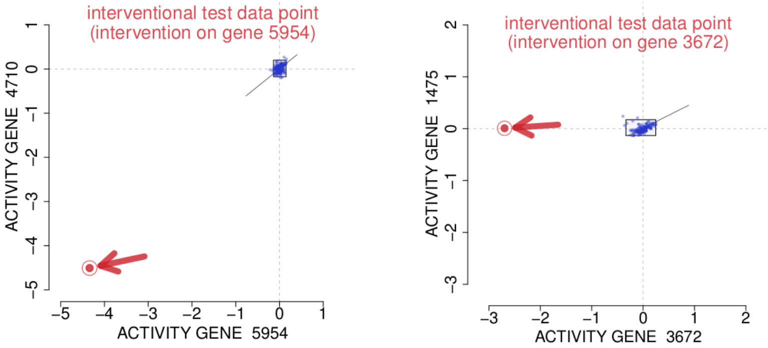
Predicting a potential outcome (synthetic data)



Predicting a potential outcome (synthetic data)



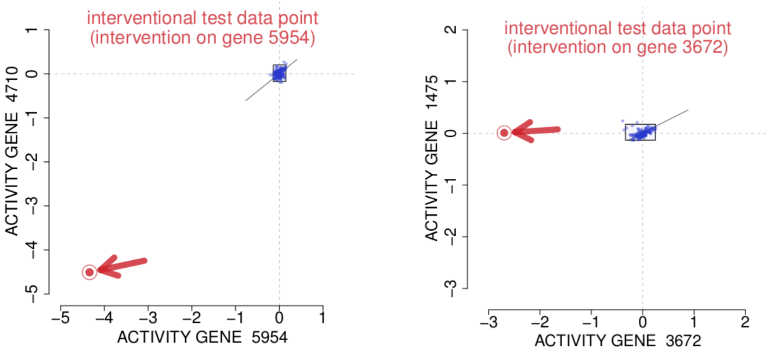
Predicting a potential outcome: real gene expression data



Challenge:

how to predict/make the correct extrapolation?

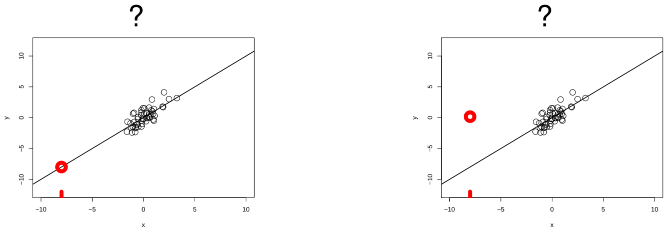
Predicting a potential outcome: real gene expression data



Challenge:

how to predict/make the correct extrapolation?

How to predict a potential outcome?



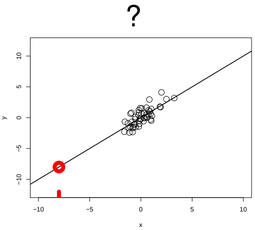
"borrow strength from other perturbations"

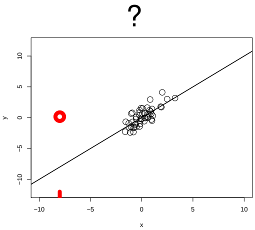
- knowing the probability distribution of the "steady state" (observational regime) is not sufficient
- it is not just regression or nonlinear deep neural nets
- the problem is the directionality! (besides the hidden confounders)

we need "some" perturbations/heterogeneities in the data

(and perturbations are often crucial for scientific discoveries)

How to predict a potential outcome?





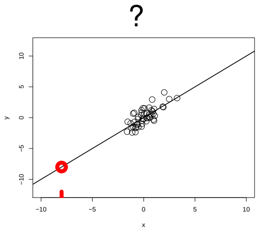
“borrow strength from other perturbations”

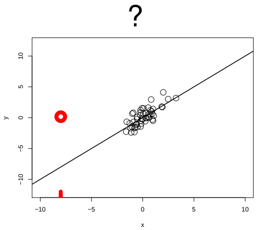
- knowing the probability distribution of the “steady state” (observational regime) is not sufficient
- it is not just regression or nonlinear deep neural nets
- the problem is the directionality! (besides the hidden confounders)

we need “some” perturbations/heterogeneities in the data

(and perturbations are often crucial for scientific discoveries)

How to predict a potential outcome?





"borrow strength from other perturbations"

- knowing the probability distribution of the "steady state" (observational regime) is not sufficient
- it is not just regression or nonlinear deep neural nets
- the problem is the directionality! (besides the hidden confounders)

we need "some" perturbations/heterogeneities in the data

(and perturbations are often crucial for scientific discoveries)

“Some references” for causal inference

- Potential outcome model
  Neyman, Holland, Rubin, Rosenbaum, ...
- Graphical and structural equation models
  Pearl, Spirtes–Glymour–Scheines, Bollen, ...
- Dawid, Robins, Richardson, Didelez, ..., Janzing–Schölkopf, Mooij, ...

we propose something “rather different”: namely
  exploit unspecific heterogeneities
or
  learn from “perturbations”
    (in contrast to “downweighting outliers/perturbations”)

Heterogeneous data: quite common in large-scale problems

data from different known observed

environments or experimental conditions or

perturbations or sub-populations $e \in \mathcal{E}$:

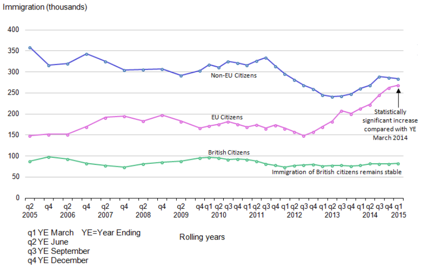
$$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

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

examples:

- data from 10 different countries
- data from different econ. scenarios (from diff. "time blocks")

immigration in the UK

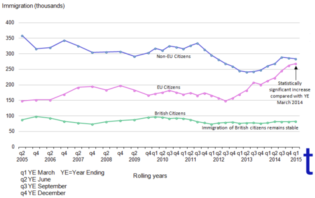


consider “many possible” but mostly non-observed environments/perturbations $\mathcal{F} \supset \underbrace{\mathcal{E}}_{\text{observed}}$

examples for $\mathcal{F}$:

- 10 countries and many other than the 10 countries
- scenarios until today and new unseen scenarios in the future

immigration in the UK



the unseen future

problem:

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

a pragmatic prediction problem:

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

for example with linear models: find

$$\text{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

and remember:

causality is predicting an answer to a "what if I do/perturb question"!

that is: prediction for new unseen scenarios/environments

a pragmatic prediction problem:

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

for example with linear models: find

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

and remember:

causality is predicting an answer to a

"what if I do/perturb question"!

that is: prediction for new unseen scenarios/environments

a pragmatic prediction problem:

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

for example with linear models: find

$$\mathrm{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

and remember:

causality is predicting an answer to a

"what if I do/perturb question"!

that is: prediction for new unseen scenarios/environments

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

for example with linear models: find

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness" and also about causality

and remember:
causality is predicting an answer to a "what if I do/perturb question"!
that is: prediction for new unseen scenarios/environments

Prediction and causality

indeed, for linear models: in a nutshell

for $\mathcal{F} = \{$all perturbations "not acting on Y directly"$\}$,
$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T\beta|^2 =$ causal parameter

that is:
causal parameter optimizes
worst case loss w.r.t. "very many" unseen ("future") scenarios

later:
we will discuss models for $\mathcal{F}$ and $\mathcal{E}$ which make these relations more precise

Prediction and causality

indeed, for linear models: in a nutshell

$$\text{for } \mathcal{F} = \{\text{all perturbations "not acting on } Y \text{ directly"}\},$$
$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2 = \text{ causal parameter}$$

that is:
causal parameter optimizes
worst case loss w.r.t. "very many" unseen ("future") scenarios

later:
we will discuss models for $\mathcal{F}$ and $\mathcal{E}$ which make these relations more precise

How to exploit heterogeneity/learn from perturbations?

a key conceptual Invariance Assumption (w.r.t. $\mathcal{E}$) :

there exists $S^* \subseteq \{1, \ldots, d\}$ 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 $\mathrm{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\begin{aligned}
\forall e \in \mathcal{E}:\ \ & Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*} \\
& \varepsilon^e \sim F_\varepsilon \text{ the same for all } e \\
& X^e \text{ has an arbitrary distribution, different across } e
\end{aligned}$$

γ^*, S^* is interesting in its own right!

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

How to exploit heterogeneity/learn from perturbations?

a key conceptual Invariance Assumption (w.r.t. $\mathcal{E}$) :

there exists $S^* \subseteq \{1, \ldots, d\}$ 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 $\mathrm{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

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

$$\varepsilon^e \sim F_\varepsilon \text{ the same for all } e$$

$$X^e \text{ has an arbitrary distribution, different across } e$$

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

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

How to exploit heterogeneity/learn from perturbations?

a key conceptual Invariance Assumption (w.r.t. $\mathcal{F}$) :

there exists $S^* \subseteq \{1, \ldots, d\}$ such that

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

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

$$\begin{aligned} \forall e \in \mathcal{F}: \ & Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*} \\ & \varepsilon^e \sim F_\varepsilon \text{ the same for all } e \\ & X^e \text{ has an arbitrary distribution, different across } e \end{aligned}$$

γ^*, S^* even more interesting
since it says something about unseen new environments!

any subset S^* of covariates satisfying the invariance ass. is

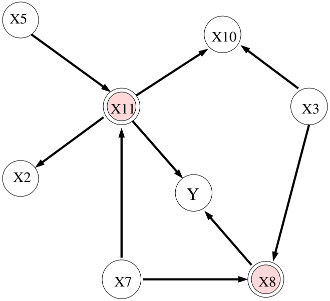
- stabilizing
- “robustifying”

the “stabilizing with invariance” will be the basis for a robustifying “procedure” (for some rather general distributional robustness)

Link to causality

mathematical formulation with structural equation models:

$$Y \leftarrow f(X_{\mathrm{pa}(Y)}, \varepsilon),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j) \ (j = 1, \ldots, p)$$
$$\varepsilon, \varepsilon_1, \ldots, \varepsilon_p \text{ independent}$$



(direct) causal variables for Y: the parental variables of Y

Link to causality

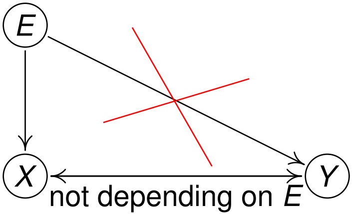
problem:
under what model for the environments/perturbations e can we have an interesting description of the invariant sets S^*?

loosely speaking: assume that the perturbations e

- do not directly act on Y
- do not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

graphical description: E is random with realizations e



Link to causality

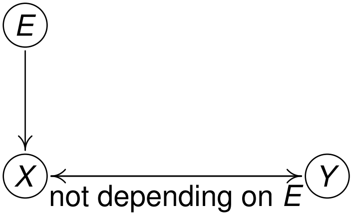
problem:
under what model for the environments/perturbations e can we have an interesting description of the invariant sets S^*?

loosely speaking: assume that the perturbations e

- do not directly act on Y
- do not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

graphical description: E is random with realizations e



Link to causality

easy to derive the following:

Proposition

- structural equation model for (Y, X);
- model for $\mathcal{F}$ of perturbations: every $e \in \mathcal{F}$
  - do not directly act on Y
  - do not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Then: the causal variables $\mathrm{pa}(Y)$ satisfy the invariance assumption with respect to $\mathcal{F}$

causal variables lead to invariance under arbitrarily strong perturbations from $\mathcal{F}$ as described above

Proposition

- structural equation model for (Y, X);
- model for $\mathcal{F}$ of perturbations: every $e \in \mathcal{F}$
  - does not directly act on Y
  - does not change the relation between X and Y

  but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Then: the causal variables $\mathrm{pa}(Y)$ satisfy the invariance assumption with respect to $\mathcal{F}$

as a consequence: for linear structural equation models

for $\mathcal{F}$ as above,

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2 = \underbrace{\beta^0_{\mathrm{pa}(Y)}}_{\text{causal parameter}}$$

if the perturbations in $\mathcal{F}$ would not be arbitrarily strong

$\leadsto$ the worst-case optimizer is different! (see later)

Proposition

- structural equation model for (Y, X);
- model for $\mathcal{F}$ of perturbations: every $e \in \mathcal{F}$
  - does not directly act on Y
  - does not change the relation between X and Y

  but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Then: the causal variables $\mathrm{pa}(Y)$ satisfy the invariance assumption with respect to $\mathcal{F}$

as a consequence: for linear structural equation models

$$\text{for } \mathcal{F} \text{ as above,}$$

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T\beta|^2 = \underbrace{\beta^0_{\mathrm{pa}(Y)}}_{\text{causal parameter}}$$

if the perturbations in $\mathcal{F}$ would not be arbitrarily strong
$\leadsto$ the worst-case optimizer is different! (see later)

A real-world example and the assumptions

Y: growth rate of the plant
X: high-dim. covariates of gene expressions

perturbations e: different gene knock-out experiments
$\rightsquigarrow$ e changes the expressions of some components of X

it's plausible that perturbations e

- do not directly act on Y $\surd$
- do not change the relation between X and Y ?

but act strongly on X (arbitrary shifts, scalings, etc.)

A real-world example and the assumptions

Y: growth rate of the plant
X: high-dim. covariates of gene expressions

perturbations e: different gene knock-out experiments
$\rightsquigarrow$ e changes the expressions of some components of X

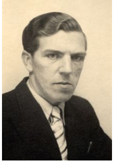
it's plausible that perturbations e

- do not directly act on Y $\surd$
- do not change the relation between X and Y ?

but act strongly on X (arbitrary shifts, scalings, etc.)

we just argued: causal variables $\Longrightarrow$ invariance

known since a long time:
Haavelmo (1943)



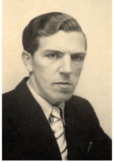
Trygve Haavelmo
Nobel Prize in Economics 1989

(...; Goldberger, 1964; Aldrich, 1989;... ; Dawid and Didelez, 2010)

Causality $\Longleftrightarrow$ Invariance

we just argued: causal variables $\Longrightarrow$ invariance

known since a long time:
Haavelmo (1943)



Trygve Haavelmo

Nobel Prize in Economics 1989

(...; Goldberger, 1964; Aldrich, 1989;... ; Dawid and Didelez, 2010)

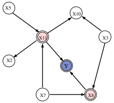
more novel: the reverse relation

causal structure, predictive robustness $\Longleftarrow$ invariance

(Peters, PB & Meinshausen, 2016
Rothenhäusler, Meinshausen, PB & Peters, 2018)

The search for invariance and causality (Peters, PB & Meinshausen, 2016)

causal structure/variables $\Longleftarrow$ invariance



severe issues of identifiability !

$\leadsto$ but can come up with a conservative procedure protecting against false positive causal selection

$$\mathbb{P}[\underbrace{\hat{S}(\mathcal{E})}_{\text{an algorithm}} \subseteq \underbrace{S_{\mathrm{causal}}}_{\mathrm{pa}(Y)}] \geq 1 - \alpha$$

which we applied to a large-scale genomic perturbation study

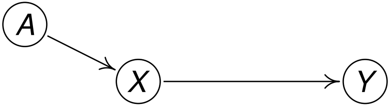
Meinshausen at al. (2016)

Anchor regression: predictive robustness and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$

(once before, we denoted it as E)



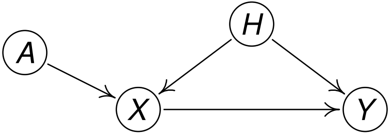
$$Y \leftarrow X^T \beta^0 + \varepsilon_Y \qquad ,$$
$$X \leftarrow A^T \alpha^0 + \varepsilon_X \qquad ,$$

Anchor regression: predictive robustness and causal regularization

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

the environments from before, denoted as e:
they are now outcomes of a variable $\underbrace{A}_{\text{anchor}}$

(once before, we denoted it as E)



$$Y \leftarrow X^T \beta^0 + \varepsilon_Y + H\delta,$$
$$X \leftarrow A^T \alpha^0 + \varepsilon_X + H\gamma,$$

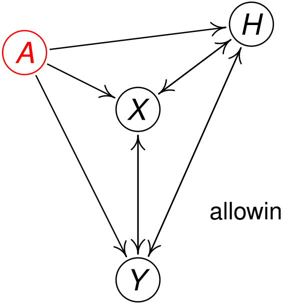
Instrumental variables regression model
(cf. Angrist, Imbens, Lemieux, Newey, Rosenbaum, Rubin,...)
hidden/latent confounders are of major concern!

Anchor regression with hidden confounders

allow that A acts on Y and H

more realistic but has been believed in the past as “ill-posed”

A is an “anchor”

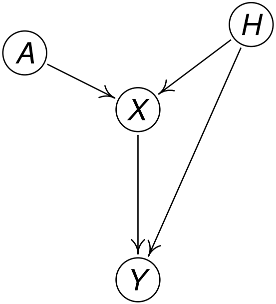


allowing also for feedback loops

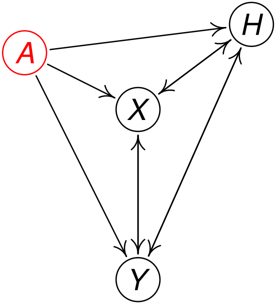
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

IV regression is a special case of anchor regression

IV regression



anchor regression



allowing also for feedback loops

allow that A acts on Y and H

$\leadsto$ there is a fundamental identifiability problem

cannot identify

the causal mechanism between $X \longleftrightarrow Y$ from data

which is the price for more realistic assumptions than IV model

... but “Causal Regularization” offers something

can still achieve “shift invariance” of residuals:
a non-trivial fact is:

$(Y - Xb)$ is “shift-invariant” $\Longleftrightarrow$ A uncorrelated with $(Y - Xb)$

thus, we want to encourage orthogonality of A with the residuals something like

$$\tilde{\beta} = \text{argmin}_b \|Y - Xb\|_2^2/n + \xi \|A^T(Y - Xb)/n\|_2^2$$

$$\tilde{\beta} = \text{argmin}_b \|Y - Xb\|_2^2/n + \xi \|A^T(Y - Xb)/n\|_2^2$$

anchor regression estimator:

$$\hat{\beta} = \text{argmin}_b \|(I - \Pi_A)(Y - Xb)\|_2^2/n + \gamma \|\Pi_A(Y - Xb)\|_2^2/n$$

$\Pi_A = A(A^T A)^{-1} A^T$ (projection onto column space of A)

- for $\gamma = 1$: ordinary least squares
- for $0 \leq \gamma < \infty$: general causal regularization

$$\tilde{\beta} = \text{argmin}_b \|Y - Xb\|_2^2/n + \xi \|A^T(Y - Xb)/n\|_2^2$$

anchor regression estimator:

$$\hat{\beta} = \text{argmin}_b \|(I - \Pi_A)(Y - Xb)\|_2^2/n + \gamma \|\Pi_A(Y - Xb)\|_2^2/n + \lambda \|b\|_1$$

$\Pi_A = A(A^T A)^{-1} A^T$ (projection onto column space of A)

- for $\gamma = 1$: least squares + ℓ_1-penalty
- for $0 \leq \gamma < \infty$: general causal regularization + ℓ_1-penalty

... there is a fundamental identifiability problem...

but causal regularization solves for

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

for a certain class of shift perturbations $\mathcal{F}$

Model for $\mathcal{F}$: (new) shifts in the (test) data

model for observed heterogeneous data ("corresponding to $\mathcal{E}$")

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

model for unobserved perturbations $\mathcal{F}$ (in test data)
shift vectors v acting on (components of) X, Y, H

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} = B \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v$$

$v \in C_\gamma \subset$ span(M), γ measuring the size of v

i.e. $v \in C_\gamma = \{v;\ v = M\delta$ for some δ with $\mathbb{E}[\delta\delta^T] \preceq \gamma\mathbb{E}[AA^T]\}$

A fundamental duality theorem

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

P_A the population projection onto A: $P_A \bullet = \mathbb{E}[\bullet|A]$

For any b

$$\begin{aligned} \max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v b|^2] &= \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right] \\ &\approx \underbrace{\|(I - \Pi_A)(Y - Xb)\|_2^2/n + \gamma \|\Pi_A(Y - Xb)\|_2^2/n}_{\text{objective function on data}} \end{aligned}$$

worst case shift interventions $\longleftrightarrow$ regularization!
the worst case L_2-loss is equal to a regularized L_2-loss

a new theory for quantitatively relating $\underbrace{\text{causality}}_{\text{interventions}}$ to robustness for any b

$$\overbrace{\max_{v \in C_\gamma} \mathbb{E}\left[\left|Y^v - X^v b\right|^2\right]}^{\text{worst case test error}}$$

$$= \underbrace{\mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]}_{\text{criterion on training population sample}}$$

$$\mathrm{argmin}_b \overbrace{\max_{v \in C_\gamma} \mathbb{E}\left[\left|Y^v - X^v b\right|^2\right]}^{\text{worst case test error}}$$

$$= \mathrm{argmin}_b \underbrace{\mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]}_{\text{criterion on training population sample}}$$

and "therefore"

$$\hat{\beta} = \mathrm{argmin}_b \|(I - \Pi_A)(Y - Xb)\|_2^2/n + \gamma \|\Pi_A(Y - Xb)\|_2^2 \ \ (+\lambda\|b\|_1)$$

protects against worst case shift intervention scenarios
and leads to

predictive stability (i.e. optimizing a worst case risk)

the word "therefore"
is justified for the high-dimensional sparse scenario

$$
\mathrm{argmin}_b \overbrace{\max_{v \in C_\gamma} \mathbb{E}\left[\left|Y^v - X^v b\right|^2\right]}^{\text{worst case test error}}
$$

$$
= \mathrm{argmin}_b \underbrace{\mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]}_{\text{criterion on training population sample}}
$$

and “therefore”

$$
\hat{\beta} = \mathrm{argmin}_b \|(I - \Pi_A)(Y - Xb)\|_2^2/n + \gamma \|\Pi_A(Y - Xb)\|_2^2 \ \ (+\lambda \|b\|_1)
$$

protects against worst case shift intervention scenarios
and leads to

predictive stability (i.e. optimizing a worst case risk)

the word “therefore”
is justified for the high-dimensional sparse scenario

Theorem (Rothenhäusler, Meinshausen, PB & Peters, 2018)
assume:

- a "causal" compatibility condition on X (weaker than the standard compatibility condition);
- (sub-) Gaussian error;
- $\dim(A) \le C < \infty$ for some C;

Then, for $R_\gamma(b) = \max_{v \in C_\gamma} \mathbb{E}|Y^v - X^v b|^2$ and any $\gamma \ge 0$:

$$R_\gamma(\hat{\beta}_\gamma) = \underbrace{\min_b R_\gamma(b)}_{\text{optimal}} + O_P(s_\gamma \sqrt{\log(d)/n}),$$

$$s_\gamma = \mathrm{supp}(\beta_\gamma), \ \beta_\gamma = \mathrm{argmin}_b R_\gamma(b)$$

if dim(A) is large: use ℓ_∞-norm causal

- good for identifiability (lots of heterogeneity) regularization
- a statistical price of $\log(|A|)$

Performance in practice

evaluate worst case risk

$$\max_v \mathbb{E}[(Y^v - X^v \hat{\beta})^2]$$

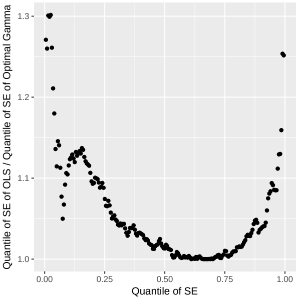
$\leadsto$ look at quantiles of $\{(Y_i - \hat{Y}_i)^2;\ i \in \text{test sample}\}$

Bike rentals in Washington DC

$n = 17'379$, $d = 4$ meteorological covariates, linear model

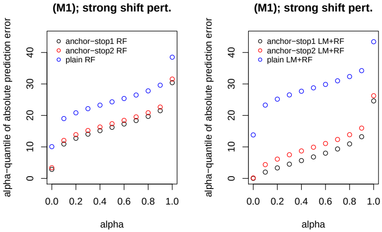
anchor = "time"

$\approx$ 15-25% gain over standard least squares



Nonlinear extensions with Random Forests

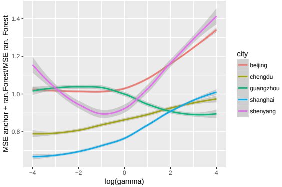
Simulations: quantiles of $\{|Y_i - \hat{Y}_i|;\ i \in \text{test sample}\}$



blue: Random Forests, black: nonlinear anchor regression with RF

Air pollution in Chinese cities

anchor: “geographical label”



some first empirical results for macro-economic predictions

- heterogeneity over different European countries
- the model is of state-space form ("state-of-the-art" model)

$\rightsquigarrow$ improvements with causal regularization (of "anchor-type")

it's a new way of thinking about "scenario robustness"!

some first empirical results for macro-economic predictions

- heterogeneity over different European countries
- the model is of state-space form ("state-of-the-art" model)

$\leadsto$ improvements with causal regularization (of "anchor-type")

it's a new way of thinking about "scenario robustness"!

It's a kind of future scenario/test sample robustness

quite different from classical statistical robustness:

- robust stats:
  – downweight outliers to "approach" the reference distr.
  – aims (primarily) for training sample robustness
- anchor/invariance: make use of/exploit the perturbations to inspect stability and hence "robustify" against against adversarial future scenarios/test samples

Connections to robust optimization

distributionally robust optimization
(Ben-Tal, El Ghaoui & Nemirovski, 2009;
e.g. Sinha, Namkoong & Duchi, 2017)

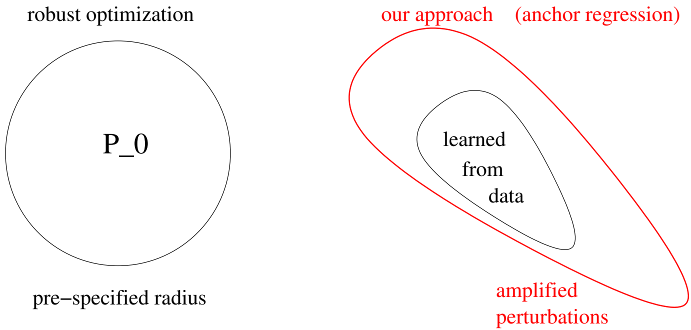
$$\text{argmin}_\beta \max_{P \in \mathcal{P}} \underbrace{\mathbb{E}_P[\ell(X,Y;\beta)]}_{\text{e.g. } \mathbb{E}_P|Y - X^T\beta|^2}$$

guarantees performance under adversarial test sample distr.

what is $\mathcal{P}$? usually

$$\mathcal{P} = \{P; \quad \underbrace{d(P, P_0)}_{\text{e.g. Wasserstein distance}} \quad \leq \rho\}$$

often $\quad P_0 = \hat{P}$ $(=$ empirical dist.$)$



anchor regression:
learn the "structure" of the class $\mathcal{F}$ from heterogeneous data in $\mathcal{E}$; and $\mathcal{F}$ is an amplification of the observed heterogeneity in $\mathcal{E}$

the class is based on a "causal-type" model $\leadsto$ has the potential for interesting interpretability

(in contrast to just having a "good metric")

Conclusions

there are surprising connections between

causality $\Longleftrightarrow$ invariance/stability $\Longleftrightarrow$ robustness

- some of them were known (e.g. Haavelmo, 1943)
- some of them are novel and
  especially interesting in the advent of large-scale data
  where perturbations/heterogeneities are unspecific

make heterogeneity, perturbations, non-stationarity your friend

(rather than your enemy)!

make heterogeneity, perturbations, non-stationarity your friend

(rather than your enemy)!

What have we done to address the "ill-posedness"?

where A is allowed to influence not only X but also Y and H

- still get predictive stability
- causality is impossible
  but $\gamma = \infty$ corresponds to invariance of residuals w.r.t. arbitrarily strong shift perturbations generated by A

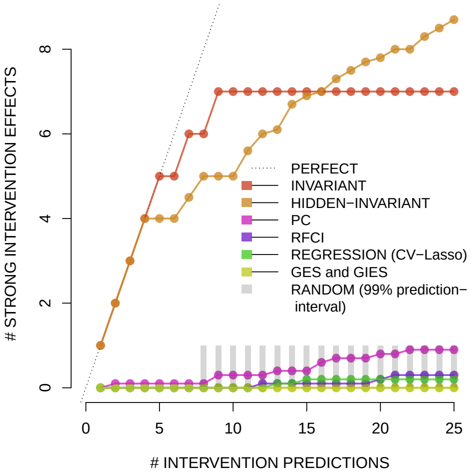
  a "diluted form of causality"
  still better than ordinary regression framework
  $\leadsto$ useful for e.g. bio-medical applications

What have we done to address the “ill-posedness”?

where A is allowed to influence not only X but also Y and H

- still get predictive stability
- causality is impossible
  but $\gamma = \infty$ corresponds to invariance of residuals w.r.t. arbitrarily strong shift perturbations generated by A

  a “diluted form of causality”
  still better than ordinary regression framework
  $\rightsquigarrow$ useful for e.g. bio-medical applications



I : invariant prediction method

H: invariant prediction with some hidden variables