Invariance and causality for robust predictions

Peter Bühlmann

Seminar für Statistik, ETH Zürich

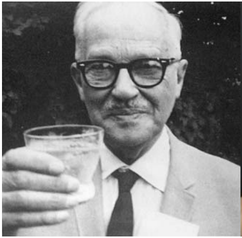
joint work with

Jonas Peters
Univ. Copenhagen

Nicolai Meinshausen
ETH Zürich

Dominik Rothenhäusler
ETH Zürich

Causality: it's (also) about predicting an answer to a "What if I do question"



Jerzy Neyman   Donald Rubin

potential outcome: what would have happened if we would have assigned a certain treatment

a main task in causality: predict a potential outcome of a certain treatment or in a certain environment based on data where this particular treatment is not observed

a main task in causality: predict a potential outcome
of a certain treatment or in a certain environment
based on data where this particular treatment is not observed

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*"
- economics: what would be the effect of a certain intervention?
  but there is no data for such a new intervention scenario

the “prediction aspect of causality” makes it

- less philosophical
- more pragmatic

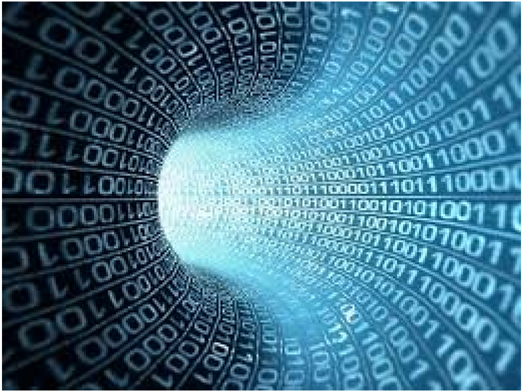
and it will allow novel notions of “robustness”

(being very different from classical robustness) there is a large body of important work on causal inference (Haavelmo, Holland, Rubin, Robins, Dawid, Pearl, Spirtes, Glymour, Scheines, Angrist, Imbens...)

"another" way of thinking and formalizing might be useful in the context of large datasets with no designed (randomized) experiments there is a large body of important work on causal inference (Haavelmo, Holland, Rubin, Robins, Dawid, Pearl, Spirtes, Glymour, Scheines, Angrist, Imbens...)

“another” way of thinking and formalizing might be useful in the context of large datasets with no designed (randomized) experiments

Causality and robustness from **Heterogeneous (large-scale) data**



we will take advantage of heterogeneity

often arising with large-scale data where

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

The setting

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

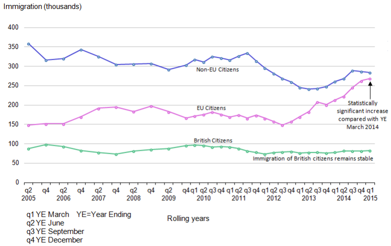
$$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

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



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

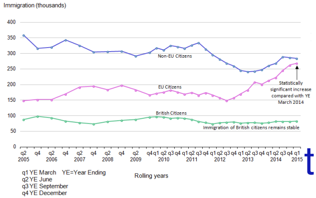
response variables Y^e, predictor variables X^e

consider "many 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
- scenarios until today and new unseen scenarios in the future

immigration in the UK



the unseen future

Prediction in heterogeneous environments

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

mostly non-observed environments $\mathcal{F} \supset \underbrace{\mathcal{E}}_{\text{observed}}$

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 $\mathcal{E}$

that is: accurate prediction which "works for new scenarios"!

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: for new (Y^e, X^e), find

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

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

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

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: for new (Y^e, X^e), find

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

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

and remember:

causality is predicting an answer to a

"what if I do/perturb question"!

that is: prediction for new unseen scenarios/environments

"equivalence": causality $\Longleftrightarrow$ prediction in heterogeneous environments

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: for new (Y^e, X^e), find

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

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

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

indeed, for linear models: in a nutshell

for $\mathcal{F} = \{$all perturbations not acting on Y directly$\}$,
$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T\beta|^2 =$ 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? for causality or “robust” prediction

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

How to exploit heterogeneity? for causality or “robust” prediction

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

Invariance: a key assumption

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:

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

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

X^e has an arbitrary distribution, different across e

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

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

Invariance: a key assumption

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:

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

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

X^e has an arbitrary distribution, different across e

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

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

$\leadsto$

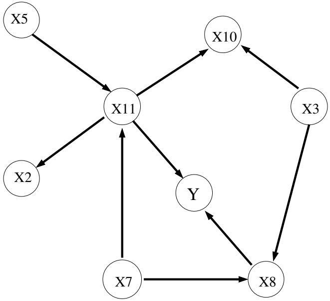
$$\gamma^*,\ S^* \text{ is even more interesting in its own right!}$$

since it says something about unseen new environments!

Link to causality

mathematical formulation with structural equation models:

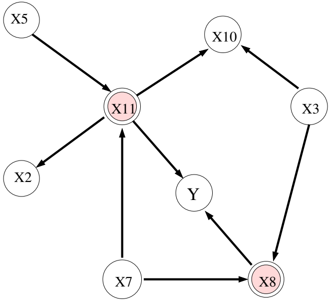
$$Y \leftarrow f(X_{\mathrm{pa}(Y)}, \varepsilon),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j) \ (j = 1, \ldots, p)$$
$$\varepsilon, \varepsilon_1, \ldots, \varepsilon_p \text{ independent}$$



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

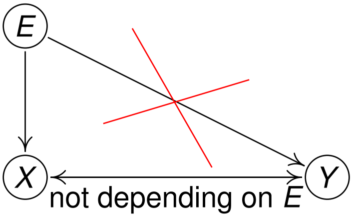
Link to causality

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

graphical description: E is random with realizations e



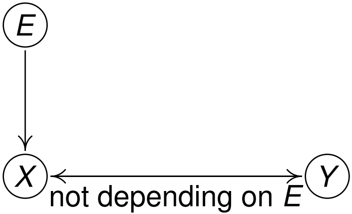
Link to causality

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
- 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 $e \in \mathcal{F}$
  - does not directly act on Y
  - does not change the relation between X and Y
  - 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
  - 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

$\rightsquigarrow$ 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
  - 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)

A real-world example and the assumptions

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

perturbations e correspond to different gene knock-out exps.
$e = 0$: observational data
$e = 1, 2, \ldots, m$: m single gene knock-out experiments

e acts in an arbitrary way on the expression of the targeted gene knock-out plus perhaps on the expression of other genes; but e is not acting directly on growth rate of plant

$\leadsto$ thus: perturbations e

- do not directly act on Y $\surd$
- do not change the relation between X and Y ?
- may act arbitrarily on X (arbitrary shifts, scalings, etc.) $\surd$

A real-world example and the assumptions

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

perturbations e correspond to different gene knock-out exps.
$e = 0$: observational data
$e = 1, 2, \ldots, m$: m single gene knock-out experiments

e acts in an arbitrary way on the expression of the targeted gene knock-out plus perhaps on the expression of other genes; but e is not acting directly on growth rate of plant

$\leadsto$ thus: perturbations e

- do not directly act on Y $\surd$
- do not change the relation between X and Y ?
- may act arbitrarily on X (arbitrary shifts, scalings, etc.) $\surd$

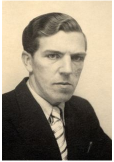
Causality $\Longleftrightarrow$ Invariance

we just argued: causal variables $\Longrightarrow$ invariance

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

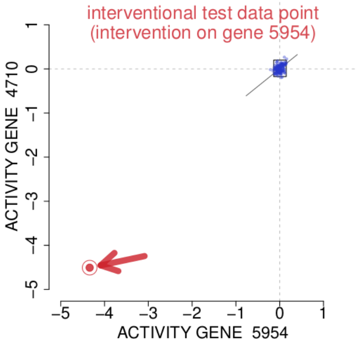
$\rightsquigarrow$ search for invariances in the data and infer causal structures
... identifiability issues! (Peters, PB & Meinshausen, 2016)

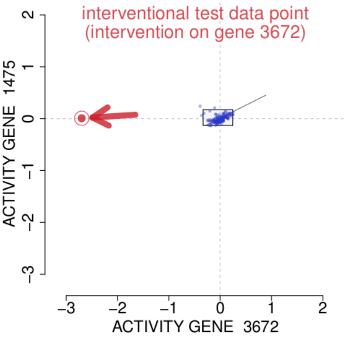
Gene knock-down perturbations

Meinshausen, Hauser, Mooij, Peters, Versteeg & PB (2016)

goal: predict gene activities (expressions) in yeast for various unobserved gene knock-down perturbations

prediction task with no data from red dots





data: gene expressions from observational data and other gene knock-down perturbations (not the ones which we want to predict)

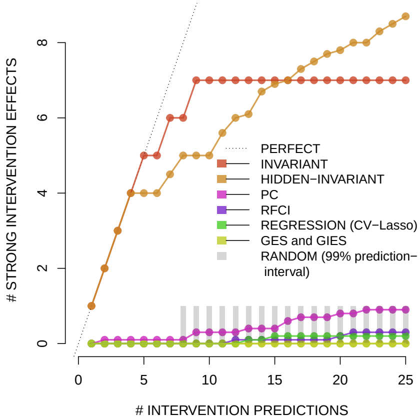
sample size: 160 observational and 1479 interventional single gene knock-down data
dimensionality: $p = 6170$ measured genes

the environments for the method (for invariance assumption): $|\mathcal{E}| = 2$, encoding “observational” and “any intervention”

put one third of the interventional samples aside (test data) and predict these interventions
validation: binarized values

strong effect (strong change): 1; otherwise: 0

predict binarized strong gene perturbations and validate with hold-out sample



I : invariant prediction method

H: invariant prediction with some hidden variables

Invariance and novel robustness

- exact invariance and corresponding causality may be often too ambitious
- the perturbations in future data might not be so strong (as in the gene knock-out example)

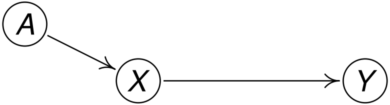
more pragmatic:
construct "best" predictions in heterogeneous settings
$\rightsquigarrow$ a novel robustness viewpoint

Anchor regression 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)



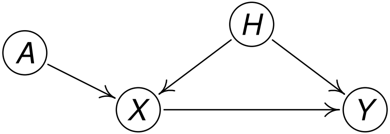
$$Y \leftarrow X^T \beta^0 + \varepsilon_Y \qquad ,$$
$$X \leftarrow A^T \alpha^0 + \varepsilon_X \qquad ,$$

Anchor regression 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)



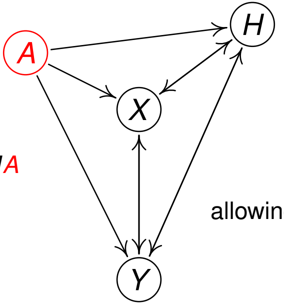
$$Y \leftarrow X^T \beta^0 + \varepsilon_Y + H\delta,$$
$$X \leftarrow A^T \alpha^0 + \varepsilon_X + H\gamma,$$

Instrumental variables regression model
(cf. Angrist, Imbens, Lemieux, Newey, Rosenbaum, Rubin,...)

hidden/latent variables are of major concern $\leadsto$ include them in the model

Anchor regression with hidden confounders

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



A is an "anchor"

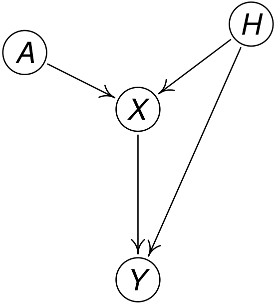
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

allowing also for feedback loops

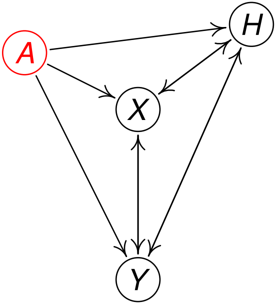
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (I - B)^{-1}(\varepsilon + MA)$$

IV regression is a special case of anchor regression

IV regression



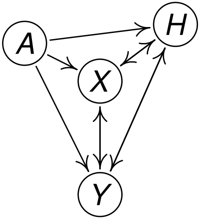
anchor regression



allowing also for feedback loops

Causal regularization

motivation: invariance assumption for residuals



when IV model does not hold
it can be shown (non-trivial!) that

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

A uncorrelated with $(Y - Xb) \iff (Y - Xb)$ is "shift-invariant"

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

something like

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

A uncorrelated with $(Y - Xb) \iff (Y - Xb)$ is "shift-invariant"

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

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

A uncorrelated with $(Y - Xb) \iff (Y - Xb)$ is "shift-invariant"

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

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 $\gamma = 0$: adjusting for heterogeneity due to A
  e.g. A are the first principal components of X capturing confounding (often used in GWAS)

A uncorrelated with $(Y - Xb) \iff (Y - Xb)$ is "shift-invariant"

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

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 $\gamma = 0$: adjusting for heterogeneity due to A
  e.g. A are the first principal components of X capturing confounding (often used in GWAS)
- for $\gamma = \infty$: two-stage least squares in IV model

A uncorrelated with $(Y - Xb) \iff (Y - Xb)$ is "shift-invariant"

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

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 $\gamma = 0$: adjusting for heterogeneity due to A
  e.g. A are the first principal components of X capturing confounding (often used in GWAS)
- for $\gamma = \infty$: two-stage least squares in IV model
- for $0 \le \gamma < \infty$: general causal regularization

A uncorrelated with $(Y - Xb) \iff (Y - Xb)$ is "shift-invariant"

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

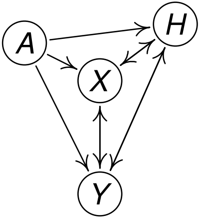
anchor regression estimator:

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

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

- for $\gamma = 1$: ordinary least squares
- for $\gamma = 0$: adjusting for heterogeneity due to A
  e.g. A are the first principal components of X capturing confounding (often used in GWAS)
- for $\gamma = \infty$: two-stage least squares in IV model
- for $0 \le \gamma < \infty$: general causal regularization + Lasso-pen.

there is a fundamental identifiability problem
since the model is more complicated than in IV regression



but causal regularization solves for

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

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

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

shift vectors v (either random or deterministic) acting on (components of) X, Y, H

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

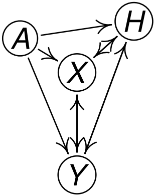
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

model for unobserved perturbations $\mathcal{F}$ (in test data)

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

$v \in \text{span}(M)$

Model for unobserved perturbations $\mathcal{F}$



consider shift interventions v acting on (X, Y, H):

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} = (I - B)^{-1}(\varepsilon + v)$$

shifts v in the $\underbrace{\text{span}(M)}_{\text{rel. to child}(A)}$, whose "strength" equals γ

$$C_\gamma = \{v;\ v = M\delta \text{ for some } \delta \text{ with } \mathbb{E}[\delta\delta^T] \preceq \gamma\mathbb{E}[AA^T]\}$$

- $\gamma = 1$: v is up to the order of MA which describes heterogeneity in the observed data
- $\gamma \gg 1$: v a strong perturbation being an amplification of the observed heterogeneity MA

Novel robustness against unobserved perturbations in $\mathcal{F}$

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

Theorem (Rothenhäusler, Meinshauen, PB & Peters, 2018)
For any b

$$\max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v b|^2] = \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]$$

worst case shift interventions $\longleftrightarrow$ regularization!

for any b

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

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

Novel robustness against unobserved perturbations in $\mathcal{F}$

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

Theorem (Rothenhäusler, Meinshauen, PB & Peters, 2018)
For any b

$$\max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v b|^2] = \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]$$

worst case shift interventions $\longleftrightarrow$ regularization!

for any b

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

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

Novel robustness against unobserved perturbations in $\mathcal{F}$

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

Theorem (Rothenhäusler, Meinshauen, PB & Peters, 2018)
For any b

$$\max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v b|^2] = \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]$$

worst case shift interventions $\longleftrightarrow$ regularization!

$$\mathrm{argmin}_b \overbrace{\max_{v \in C_\gamma} \mathbb{E}\left[\left|Y^v - X^v b\right|^2\right]}^{\text{worst case test error}}$$
$$= \mathrm{argmin}_b \underbrace{\mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - Xb)\right|^2\right] + \gamma \mathbb{E}\left[\left|P_A(Y - Xb)\right|^2\right]}_{\text{criterion on training population sample}}$$

and “therefore”

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

protects against worst case shift intervention scenarios
and leads to predictive stability

Justification of $\hat{\beta}$ in the high-dimensional 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 $\gamma \ge 0$:

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

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

Bike rentals: robust prediction

data from UCI machine learning repository
hourly counts of bike rentals between 2011 and 2012 of the "Capital Bikeshare" in Washington D.C.
sample size $n = 17'379$

goal: predict bike rentals based on the $d = 4$ covariates
temperature, feeling temperature, humidity, windspeed

use discrete anchor variable = "time":
block of consecutive time points from every day is one level

results are adjusted for hour, working day, weekday, holiday

want to evaluate worst case risk

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

worst case risk

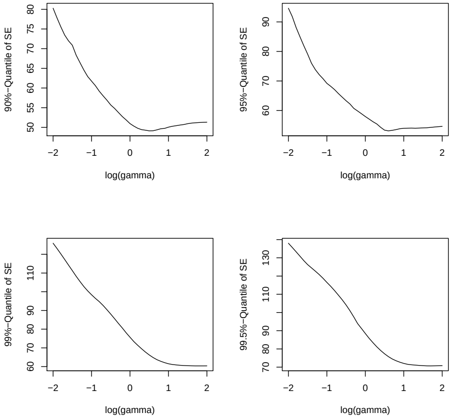
$$\max_{v} \mathbb{E}[(Y^v - X^v\hat{\beta})^2]$$

can show (under the model assumptions) that this corresponds to quantiles of $\mathbb{E}[(Y - X\hat{\beta})^2|A]$:

$$\max_{v \in C_\gamma} \mathbb{E}[(Y^v - X^v\hat{\beta})^2] = \alpha_\gamma - \text{quantile of } \mathbb{E}[(Y - X\hat{\beta})^2|A]$$

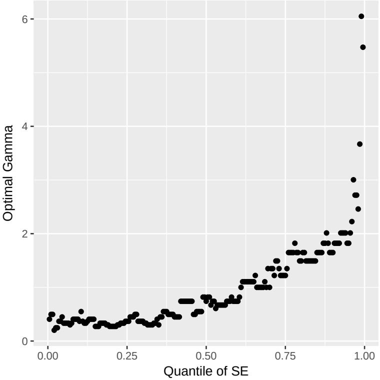
$$\gamma \text{ large} \iff \alpha = \alpha_\gamma \text{ large}$$

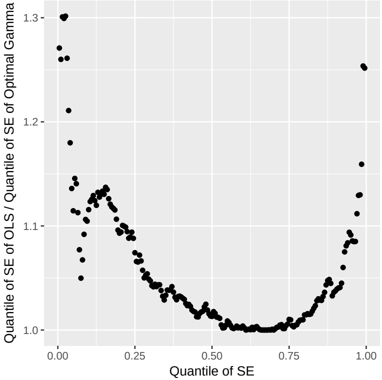
thus:
for perturbations with large v we have to look at high quantiles



large γ lead to better cross-validated performance for high quantiles of $\mathbb{E}[(Y - X\hat{\beta})^2|A]$ corresponding to worst case risk

$$\max_{v \in C_\gamma} \mathbb{E}[(Y^v - X^v\hat{\beta})^2]$$ for large class C_γ





large γ good for high quantiles of CV squared error; and vice-versa

up to 25% performance gain for high quantiles of CV squared error

It's simply transformed variables

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

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

build

$$\tilde{X} = (I - \Pi_A)X + \sqrt{\gamma}\Pi_A X = (I - (1 - \sqrt{\gamma})\Pi_A)X$$
$$\tilde{Y} = (I - \Pi_A)Y + \sqrt{\gamma}\Pi_A Y = (I - (1 - \sqrt{\gamma})\Pi_A)Y$$

then: OLS/Lasso on $(\tilde{Y}, \tilde{X})$ leads to unpenalized /ℓ_1-norm penalized anchor regression

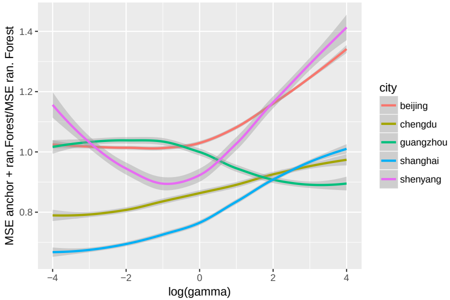
can also use nonlinear techniques with $\tilde{Y}, \tilde{X}$ as input
$\rightsquigarrow$ work in progress

Random Forests with $\tilde{Y}, \tilde{X}$ as input

Air pollution in Chinese cities

sample size $n \approx 290'000$, $p = 10$ covariables, 5 Chinese cities
anchors: the 5 different cities (different environments)

goal: predict air pollution of one city based on others



small values of γ are good $\rightsquigarrow$ the unseen perturbations are "orthogonal" to the observed heterogeneity in the data

perhaps these ideas are also
useful in the context of forecasting in economics

(e.g. unemployment, GDP,... :
currently a master thesis in collaboration with the KOF Swiss Economic Institute, ETH Zurich)

Conclusions

Invariance and Stability $\longleftrightarrow$ Causality
causal components remain the same for
different sub-populations, experimental settings or “regimes”

Shift perturbations $\longleftrightarrow$ Causal regularization
$\leadsto$ predictive stability, robustness

$\leadsto$ there are interesting and perhaps “surprising” connections between causality and predictive stability/robustness make heterogeneity or non-stationarity your friend

(rather than your enemy)!

make heterogeneity or non-stationarity your friend

(rather than your enemy)!

more on quantiles of CV squared error performance