Causal Regularization

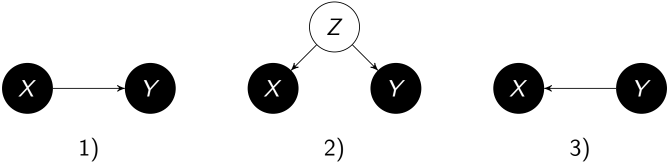
Dominik Janzing

AWS Causality Team, Amazon Research Tübingen, Germany

Oct 2019

Reichenbach's principle of common cause (1956)

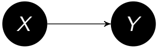
If two variables X and Y are statistically dependent then either



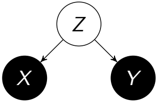
the cases are not exclusive

Reichenbach: The direction of time, 1956

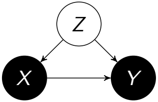
Possible scenarios if $Y \to X$ can be excluded (time order)



purely causal



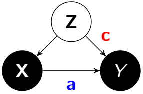
purely confounded



confounded causal relation

- $p(y|do(x)) = p(y|x)$ only for the first case
- ambitious goal: try to infer $p(y|do(x))$ from $P_{X,Y}$
- here: $\mathbf{X}, \mathbf{Z}$ high-dimensional, while Y is one-dimensional

Statistical versus causal model in a linear scenario



- **causal model:** $Y = \mathbf{X}\mathbf{a} + \mathbf{Z}\mathbf{c}$
  (in interventions $\mathbf{Z}\mathbf{c}$ is an independent noise term)

- **statistical model:** $Y = \mathbf{X}\hat{\mathbf{a}} + E \qquad E \perp\!\!\!\perp \mathbf{X}$
  with OLS regression vector

$$\hat{\mathbf{a}} := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y} = \mathbf{a} + \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XZ}}\mathbf{c}$$

$\mathbf{a}, \hat{\mathbf{a}}$ correspond to $p(y|do(\mathbf{x}))$ and $p(y|\mathbf{x})$ respectively[1]

[1] if distributions are Gaussian and or linear prediction is considered

Idea:

Regularization helps against overfitting finite data,

one should also regularize in the infinite sample limit

to obtain *causal* models instead of *statistical* models

Not too surprising because...

- regularization helps to generalize
- models that generalize across different environments are often causal models or at least causal models work better (papers of Schölkopf, Peters, Zhang, Bühlmann, Meinshausen,...)
- hence regularization helps in finding causal models

However...

- I just have *one* environment
- want to ‘generalize’ from statistical model to causal model
- not really ‘generalization’, but still possible subject to assumptions
- try to find a setting where analogy between overfitting and confounding gets as tight as possible and
- where exactly the same regularization helps against both

Standard supervised prediction problem

infer real-valued target variable Y from d-dimensional predictor variable $\mathbf{X} := (X_1, \dots, X_d)$,

- **empirical data:** $d \times n$ data matrix $\hat{\mathbf{X}}$ and vector $\hat{Y} \in \mathbb{R}^n$
- **goal:** infer y_{n+1} from $\mathbf{x}_{n+1}$

Ordinary least squares regression

- **assumption:** linear statistical model

$$Y = \mathbf{X}\mathbf{a} + E \quad \text{with } E \perp\!\!\!\perp \mathbf{X}$$

- **inference:** infer $\mathbf{a}$ via

$$\hat{\mathbf{a}}_0 := \widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}Y}} = \widehat{\Sigma_{\mathbf{XX}}}^{-1}(\widehat{\Sigma_{\mathbf{XX}}}\mathbf{a}+\widehat{\Sigma_{\mathbf{X}E}}) = \mathbf{a}+ \underbrace{\widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}E}}}_{\text{overfitting error}} .$$

- **overfitting problem:** empirical correlations between $\mathbf{X}$ and E.

Regularization

- **Ridge regression:** L^2 norm as penalizing term

$$\hat{\mathbf{a}}_\lambda := \mathrm{argmin}_{\mathbf{a}'}\{\|\hat{\mathbf{X}}\mathbf{a}' - \hat{Y}\|^2 + \lambda\|\mathbf{a}'\|^2\}$$

- **Lasso regression:** L^1 norm as penalizing term

$$\hat{\mathbf{a}}_\lambda := \mathrm{argmin}_{\mathbf{a}'}\{\|\hat{\mathbf{X}}\mathbf{a}' - \hat{Y}\|^2 + \lambda\|\mathbf{a}'\|_1\}.$$

(choose λ via cross validation)

Bayesian view on Ridge and Lasso

- Gaussian / Laplacian prior on $\mathbf{a}$: $\mathcal{N}(0, \tau^2 \mathbf{I})$ or $\mathrm{Laplace}(0, \tau^2 \mathbf{I})$
- Gaussian noise $E \sim \mathcal{N}(0, \sigma_E^2)$
- Ridge / Lasso maximize posterior with $\lambda := \sigma_E^2 / \tau^2$:

$$\log p(\mathbf{a} | \hat{\mathbf{X}}, \hat{Y}) \stackrel{+}{=} -\|\hat{\mathbf{X}}\mathbf{a} - \hat{Y}\|^2 - \frac{\sigma_E^2}{\tau^2} \|\mathbf{a}\|_{(1)}^{(2)}.$$

- rewrite in terms of covariances

$$\log p(\mathbf{a} | \widehat{\Sigma_{\mathbf{XX}}}, \widehat{\Sigma_{\mathbf{X}Y}}) \stackrel{+}{=} (\mathbf{a} - \hat{\mathbf{a}})^T \widehat{\Sigma_{\mathbf{XX}}} (\mathbf{a} - \hat{\mathbf{a}}) + \lambda \|\mathbf{a}\|_{(1)}^{(2)}$$

  with $\hat{\mathbf{a}} := \widehat{\Sigma_{\mathbf{XX}}}^{-1} \widehat{\Sigma_{\mathbf{X}Y}}$

- define population Ridge and Lasso by replacing $\widehat{\Sigma_{\mathbf{XX}}}, \widehat{\Sigma_{\mathbf{X}Y}}$ with $\Sigma_{\mathbf{XX}}, \Sigma_{\mathbf{X}Y}$

Causal regression problem in the population limit

- **assumption:** linear causal model

$$Y = \mathbf{X}\mathbf{a} + E \quad \text{with } E \not\perp \mathbf{X}$$

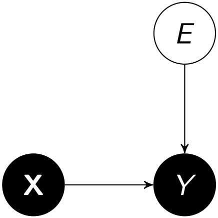
- **inference:** infer $\mathbf{a}$ via

$$\hat{\mathbf{a}}_0 := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y} = \Sigma_{\mathbf{XX}}^{-1}(\Sigma_{\mathbf{XX}}\mathbf{a} + \Sigma_{\mathbf{X}E}) = \mathbf{a} + \underbrace{\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}E}}_{\text{confounding error}} .$$

- **confounder problem:** correlations between $\mathbf{X}$ and E

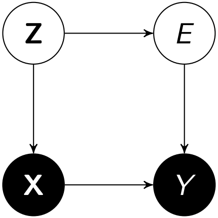
whether it's overfitting or confounding, both kinds of errors are due to correlations between $\mathbf{X}$ and E

Analogy between overfitting and confounding



$= \mathbf{X}\mathbf{a} + E$

$\widehat{\Sigma_{\mathbf{X}E}} \neq 0$ due to finite sampling



$= \mathbf{X}\mathbf{a} + E$

$\Sigma_{\mathbf{X}E} \neq 0$ due to confounding

does a regression algorithm care about why $\widehat{\Sigma_{\mathbf{X}E}} \neq 0$?

Getting the analogy even closer

- **observation:** $\widehat{\Sigma_{\mathbf{X}E}} \sim \mathcal{N}(0, \widehat{\Sigma_{\mathbf{XX}}} \frac{\sigma_E^2}{n})$

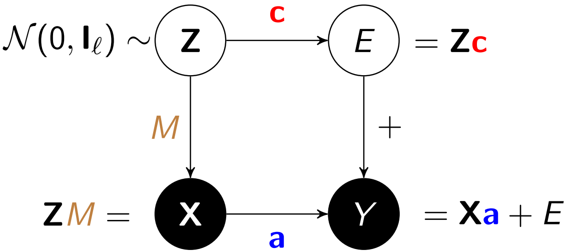
- **goal:** construct a generating model for confounders that generates $\Sigma_{\mathbf{X}E}$ according to $\mathcal{N}(0, \Sigma_{\mathbf{XX}}\gamma)$ for some parameter γ

- **conclusion:** then population versions of Ridge and Lasso maximize posterior $p(\mathbf{a}|\Sigma_{\mathbf{XX}}, \Sigma_{\mathbf{X}Y})$

Independent sources model of confounding

Janzing & Schölkopf, ICML 2018

- compose $\mathbf{X}$ from $\mathbf{Z}$ by a fixed $d \times \ell$ mixing matrix M
- compose E from $\mathbf{Z}$ by a random mixing vector $\mathbf{c} \sim \mathcal{N}(0, \mathbf{I}_\ell \frac{\sigma_c^2}{\ell})$



- then $\Sigma_{\mathbf{X}E} = M^T \mathbf{c} \sim \mathcal{N}(0, \Sigma_{\mathbf{XX}} \frac{\sigma_c^2}{\ell})$ as desired!
- number of sources replaces sample size
- confounding parameter σ_c replaces noise level

Problem

- we don't know number of sources

- we don't know confounding parameter

- how should be choose the the regularization term?
  (cross validation would need different environments)

Estimating strength of confounding DJ & Schölkopf, ICML 2018

- error vector $\mathbf{a}_e := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}E}$ is strongly concentrated in the low eigenvalue subspace of $\Sigma_{\mathbf{XX}}$
- compute ordinary least squares regression $\hat{\mathbf{a}} = \mathbf{a} + \mathbf{a}_e$
- the larger $\mathbf{a}_e$ the more $\hat{\mathbf{a}}$ concentrates in low eigenvalue subspace of $\Sigma_{\mathbf{XX}}$ (assuming an isotropic prior for $\mathbf{a}$)
- provides a rough estimation of the optimal regularization parameter (performed not too bad for dimensions $10-30$, also for two real data sets with one-dimensional confounder)

Algorithm ConCorr

1. **Input:** i.i.d. samples from $P(\mathbf{X}, Y)$.
2. Compute OLS regression vector $\hat{\mathbf{a}} := \widehat{\Sigma_{\mathbf{XX}}}^{-1}\widehat{\Sigma_{\mathbf{X}Y}}$
3. Estimate length of causal vector $\mathbf{a}$ from the estimated confounding strength
4. find λ such that the squared length of $\hat{\mathbf{a}}_\lambda^{\mathtt{ridge/lasso}}$ coincides with estimated length

Simulation Results

Generate entries of $M, \mathbf{a}, \mathbf{c}$ from $\mathcal{N}(0,1), \mathcal{N}(0,\sigma_a^2), \mathcal{N}(0,\sigma_c^2)$ after σ_a, σ_c are uniformly drawn from $[0,1]$.

$$\mathbf{X} = M\mathbf{Z} \qquad Y = \mathbf{X}\mathbf{a} + \mathbf{Z}\mathbf{c} + E.$$

Define relative squared error $RSE := \frac{\|\hat{\mathbf{a}}_\lambda - \mathbf{a}\|^2}{\|\hat{\mathbf{a}}_\lambda - \mathbf{a}\|^2 + \|\mathbf{a}\|^2}$

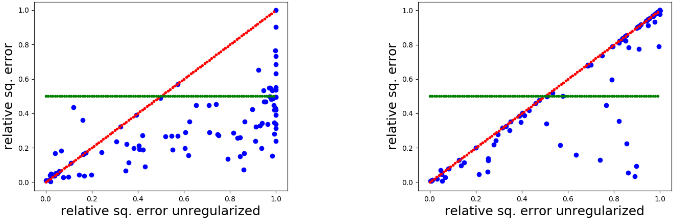
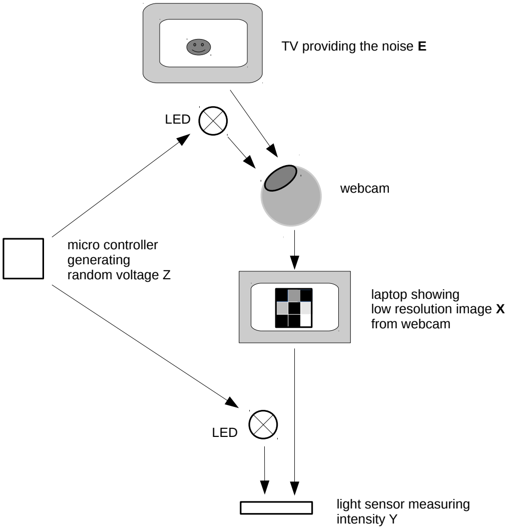
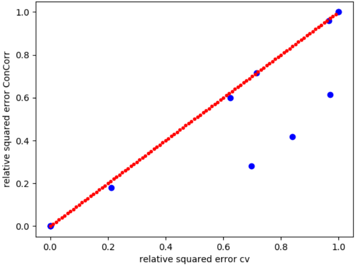


Figure: Left: Lasso with `ConCorr`. right: Lasso with cross validation

Optical device with known confounder



Results



In 3 out of 12 cases `Concorr` yielded a regression vector closer to the true one. It never got worse than unregularized regression.

Exp. with partially known confounding: taste of wine

(UCI machine learning repository)

- **cause X:** 11 measured ingredients of wine
- **effect** Y**:** taste (response of human subjects)
- **dominant cause:** X_{11} alcohol
- **generate confounding:** drop X_{11}

`Concorr` reduced the confounding error roughly by the factor $1/2$.

More general results? Is there a causal learning theory?

- are statistical relations more likely to be causal if they can be described by functions from a small class?

- could there be a learning theory for ‘generalizing’ from *statistical* relations to *causal* relations?

(without knowing the confounder)

Goal of causal learning theory that I have in mind

Infer expected interventional loss

$$\mathbb{E}_{do(X)}[(Y - f(\mathbf{X}))^2] := \int (y - f(x))^2 p(y|do(x))p(x)dydx$$

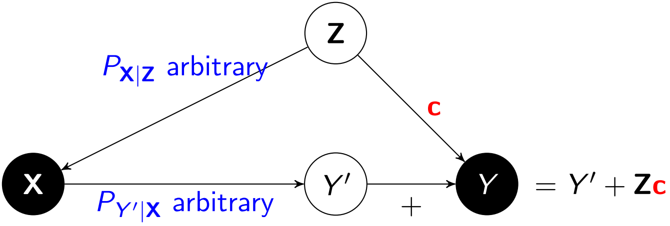
from expected statistical loss

$$\mathbb{E}[(Y - f(\mathbf{X}))^2] := \int (y - f(x))^2 p(y|x)p(x)dydx.$$

only possible subject to an appropriate confounder model!

Confounder model with linear shift

- ℓ independent sources $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_\ell)$
- $\mathbf{Z}$ causes shift $\mathbf{Z}\mathbf{c}$ of Y with $\mathbf{c} \sim \mathcal{N}(0, \sigma_c^2 \mathbf{I}_\ell)$



Causal generalization bound

For all $f \in \mathcal{F}$

$$\mathbb{E}_{do(X)}[(Y - f(\mathbf{X}))^2] \leq \mathbb{E}[(Y - f(\mathbf{X}))^2] + \epsilon,$$

holds with probability δ,

where ϵ depends on

- δ
- some appropriate capacity measure $C(\mathcal{F})$
- the confounding parameter σ_c
- the number ℓ of sources

Confounder dependent capacity for $\mathcal{F}$

Definition (correlation dimension)

Let $\mathcal{F}$ be some class of functions $f : \mathbb{R}^d \to \mathbb{R}$. Given the distribution $P_{\mathbf{X},\mathbf{Z}}$, the correlation dimension d_{corr} of $\mathcal{F}$ is the dimension of the span of

$$\{\Sigma_{f(\mathbf{X})\mathbf{Z}} \,|f \in \mathcal{F}\}.$$

Examples:

- if $\mathcal{F}$ is a function space of dimension $d_{\mathcal{F}}$ then $d_{\mathrm{corr}} \leq d_{\mathcal{F}}$
- if $\mathcal{F}$ is the space of linear functions then $d_{\mathrm{corr}} \leq \mathrm{rank}(\Sigma_{\mathbf{XZ}})$

Causal generalization bound

Define $g(\mathbf{x}) := \mathbb{E}[Y'|\mathbf{x}]$.

Let $\mathcal{F}$ have correlation dimension d_{corr} and satisfy the bound $\mathbb{E}[(f(\mathbf{X}) - g(\mathbf{X})^2] \leq b$ for all $f \in \mathcal{F}$ and let $\mathbf{c}$ be drawn from the Haar measure on the unit sphere with radius $\sqrt{V}$. Then, for any $\beta > 1$,

$$\mathbb{E}_{do(\mathbf{x})}[(Y - f(\mathbf{X})^2] \leq \mathbb{E}[(Y - f(\mathbf{X}))^2] + b \cdot \sqrt{V \cdot \beta \cdot \frac{d_{\mathrm{corr}} + 1}{\ell}},$$

holds uniformly for all $f \in \mathcal{F}$ with probability $e^{n(1-\beta+\ln\beta)/2}$.

Idea / conclusions

- for $\mathcal{F}$ with low correlation dimension, any $f \in \mathcal{F}$ will typically provide a bad *statistical* model because the confounding term is too complex to be learned

- but then we can be more sure that this model also contains *causal* truth

- regularize more than the sample size suggests (unless better methods apply)

slightly increases the chances of getting causal information.

Thank you for your attention!