Causal Dantzig

fast inference in linear structural equation models

Dominik Rothenhäusler, Peter Bühlmann and Nicolai Meinshausen

10th July, Luminy

Data: Let (X_i, Y_i) for $i = 1, \ldots, n$ be iid samples of predictor $X \in \mathbb{R}^p$ and response Y with joint distribution F. Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and reponse vector $\mathbf{Y} \in \mathbb{R}^n$.

Goal of regression: find 'optimal prediction coefficients'

$$\text{argmin}_{\beta \in \mathbb{R}^p} \; E_F\left[(Y - X'\beta)^2\right],$$

where the expectation is with respect to the sampling distribution F.

Data: Let (X_i, Y_i) for $i = 1, \ldots, n$ be iid samples of predictor $X \in \mathbb{R}^p$ and response Y with joint distribution F. Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and reponse vector $\mathbf{Y} \in \mathbb{R}^n$.

Goal of regression: find 'optimal prediction coefficients'

$$\text{argmin}_{\beta \in \mathbb{R}^p} \; E_F\left[(Y - X'\beta)^2\right],$$

where the expectation is with respect to the sampling distribution F.

Goal of 'causal inference': find a coefficient vector with optimal predictive accuracy over a class of distributions $\mathcal{F}$ (with $F \in \mathcal{F}$):

$$\text{argmin}_{\beta \in \mathbb{R}^p} \max_{F' \in \mathcal{F}} E_{F'}\left[(Y - X'\beta)^2\right].$$

1. Prediction valid for $F = F_{obs}$ versus valid for a whole class $\mathcal{F}$ of distributions

Green tea may hold the key to long life

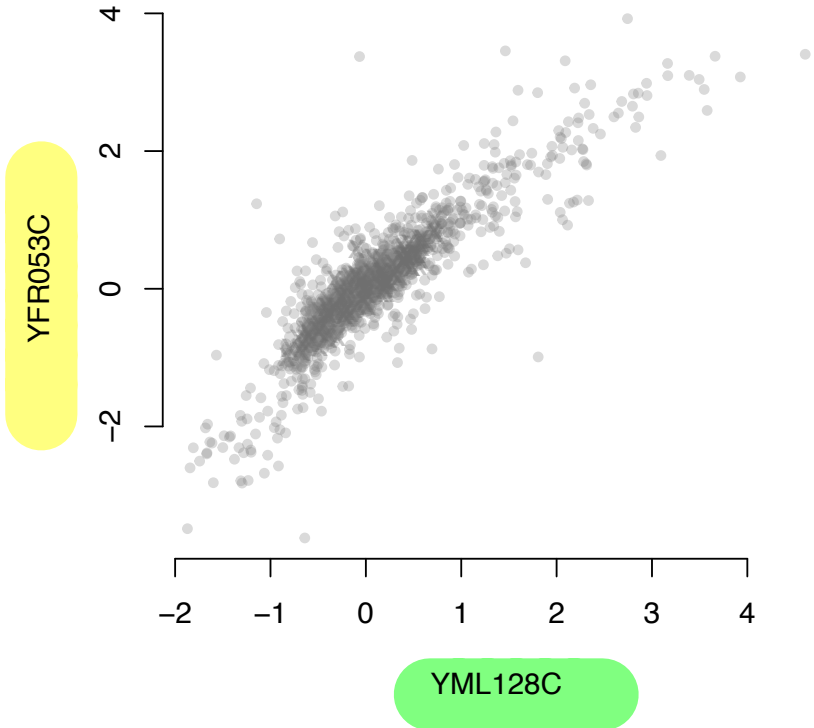
By Roger Highfield, Science Editor
12:01AM BST 13 Sep 2006

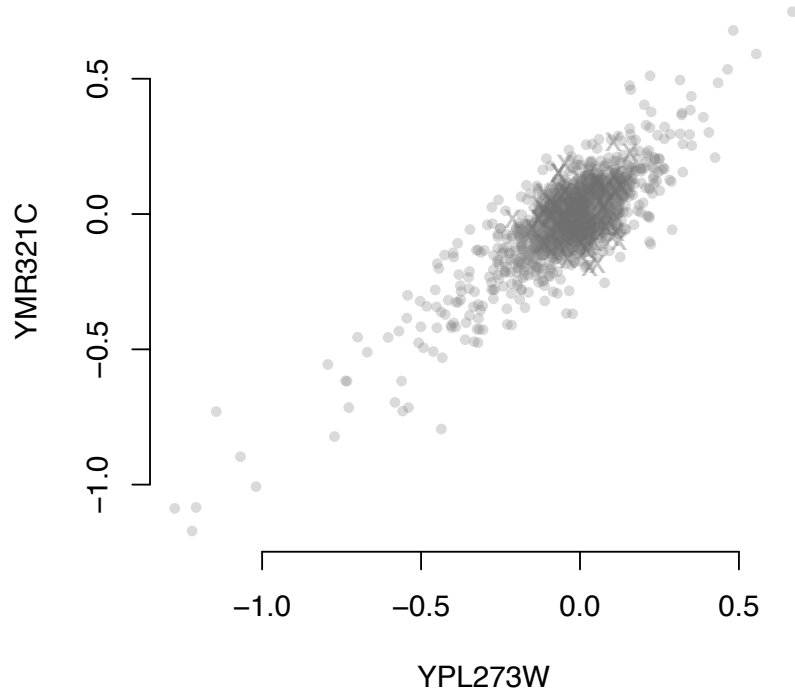
Green tea can make you live longer, with women getting a greater health benefit from the drink than men.

People who consumed higher amounts of green tea had a lower risk of death due to all causes, according to the study of more than 40,000 adults published today.

2. Computational and statistical aspects of 'large p' and variable selection

Kemmeren data: gene activities for around 5000 genes in yeast from observational ($n_{obs} = 160$) and (vaguely specified) intervention data ($n_{int} \approx 1400$).





What happens to YFR053C if we knock out/delete gene YML128C?
What happens to YMR321C if we knock out/delete gene YPL273W?

Variable selection for regression: find coefficients $\beta \in \mathbb{R}^p$ with $S = \{k : \beta_k \neq 0\} \subseteq \{1, \ldots, p\}$ such that

$$\text{Res}(\beta) \perp\!\!\!\perp X_{S^c},$$

where $\text{Res}(\beta) = Y - X\beta$.

Variable selection for regression: find coefficients $\beta \in \mathbb{R}^p$ with $S = \{k : \beta_k \neq 0\} \subseteq \{1, \ldots, p\}$ such that

$$\text{Res}(\beta) \perp\!\!\!\perp X_{S^c},$$

where $\text{Res}(\beta) = Y - X\beta$.

Variable selection for causality: find (possibly sparse) coefficients $\beta \in \mathbb{R}^p$ such that

$$\text{Res}(\beta) \quad \stackrel{d}{=} \quad \text{Res}'(\beta),$$

if left hand side is under observational distribution F_{obs} for (Y, X) and right hand side under a distribution F' for (Y, X) where we intervene on X in some way.

Observational distribution – Regression

Lasso regression for 'large p' (Tibshirani 96):

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1.$$

Lasso regression for 'large p' (Tibshirani 96):

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

KKT conditions for solution:

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \lambda \mathbf{s},$$

where $\mathbf{s} \in \partial\|\beta\|_1$, that is

$$\mathbf{s}_k = \begin{cases} 1 & \text{if } \beta_k > 0 \\ [-1, 1] & \text{if } \beta_k = 0 \\ -1 & \text{if } \beta_k < 0 \end{cases}$$

Lasso regression for 'large p' (Tibshirani 96):

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

KKT conditions for solution:

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \lambda \mathbf{s},$$

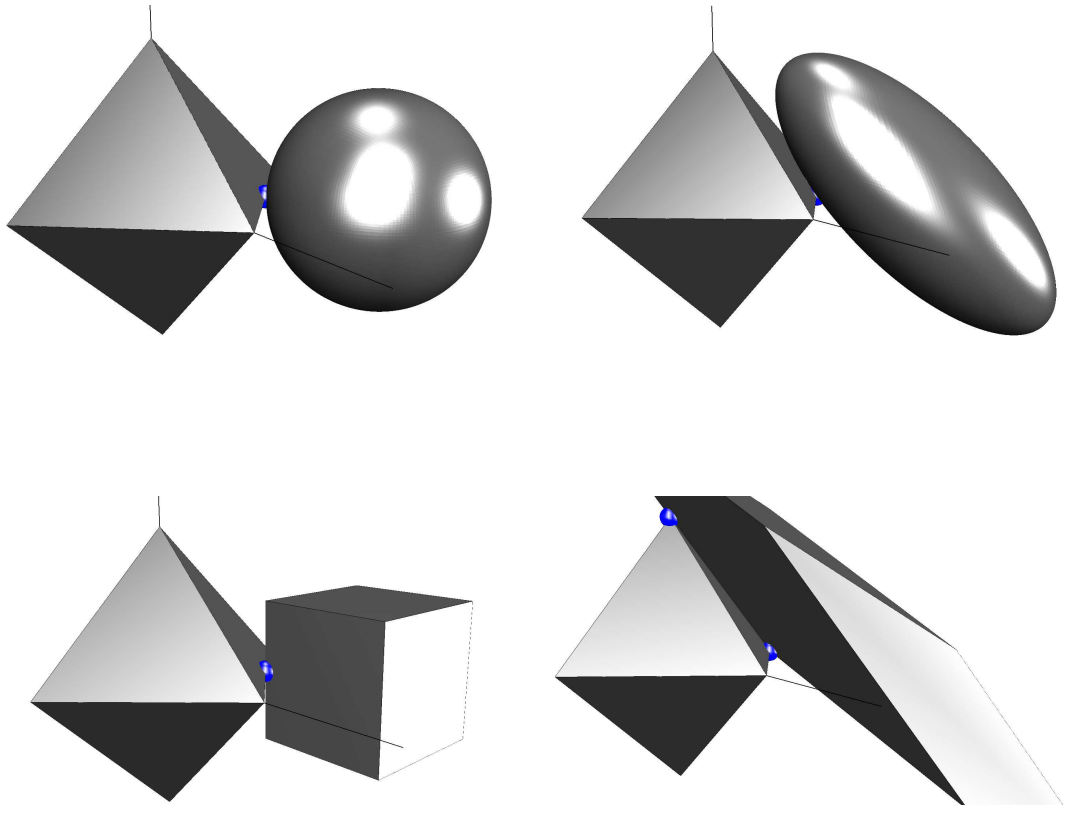
where $\mathbf{s} \in \partial\|\beta\|_1$, that is

$$\mathbf{s}_k = \begin{cases} 1 & \text{if } \beta_k > 0 \\ [-1, 1] & \text{if } \beta_k = 0 \\ -1 & \text{if } \beta_k < 0 \end{cases}$$

or $\|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty \leq \lambda$ (and some sign constraints).

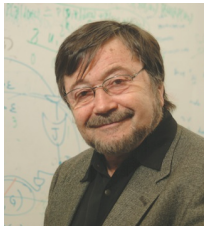
Dantzig selector (Candes and Tao, '07):

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty \leq \lambda.$$



Analysis in Candes and Tao ('07), Bickel et al. ('09) and Ye and Zhang ('10)

Interventions – Causality



Donald Rubin

Judea Pearl

Phil Dawid

Thomas Richardson

James Robins

1. How to describe a revelant class $\mathcal{F}$ of distributions under interventions?
2. Is 'large p' inference possible?

Potential outcome model:

We can only observe one of

$$Y_{\text{treatment}}, \quad Y_{\text{control}}.$$

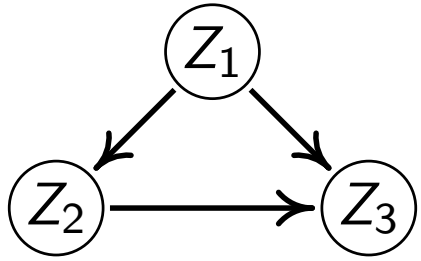
Goal: infer average causal effect $E(Y_{\text{treatment}} - Y_{\text{control}})$

Nayman, 1923, Wilk, 1955, Reichenbach, 1956; Suppes, 1970; Rubin, 1974; Dawid, 1979; Holland, 1986,...

Structural equation models:

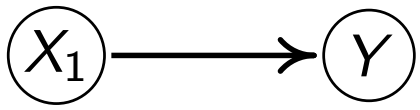
Let $Z = (Y, X)$. A linear structural equation model for $(Z_1, ..., Z_{p+1})$ is of the form (Bollen et al. 89, Robins et al. 00, Pearl 09)

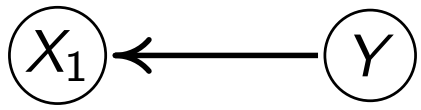
$$Z_k \leftarrow \sum_{k' \neq k} A_{k,k'} Z_{k'} + \eta_k, \qquad \text{for } k = 1, \ldots, p+1$$

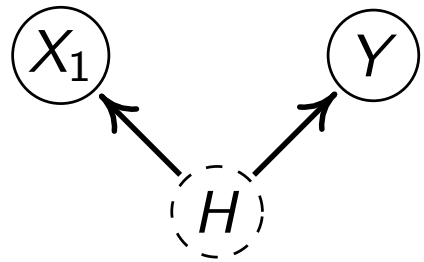


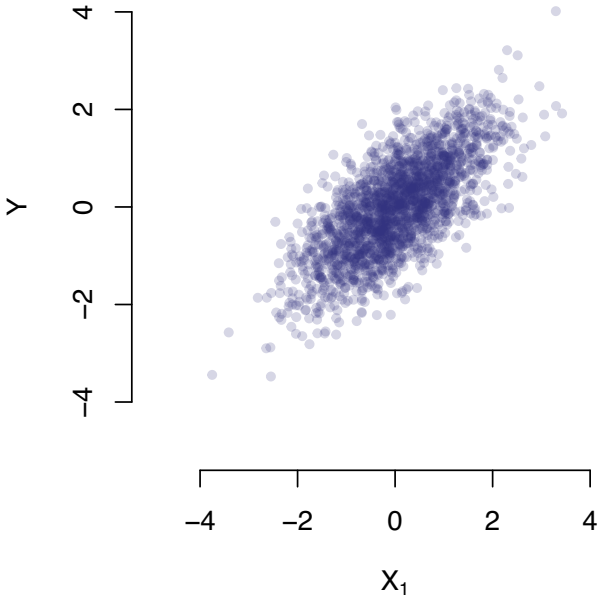
Parents of variable Z_k are $\text{parents}(Z_k) = \{k' : A_{k,k'} \neq 0\}$.

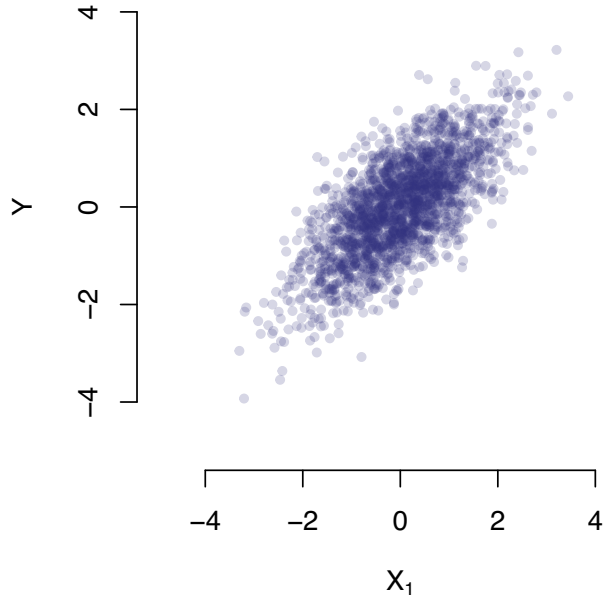
Many SEM generate the same observational distribution

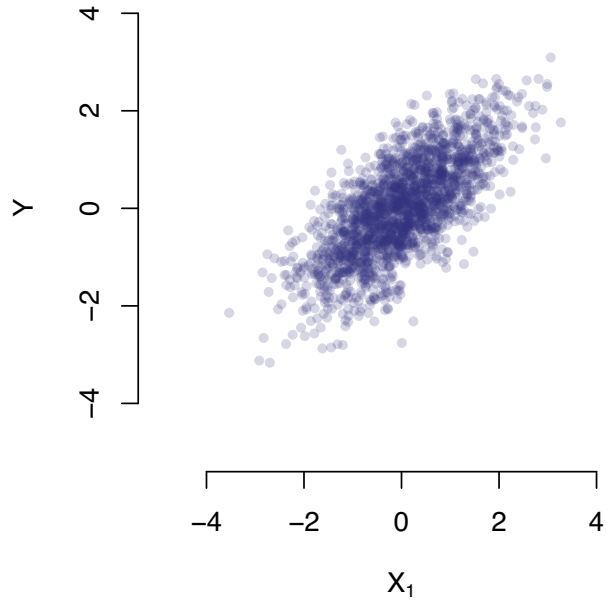












Intervention on Z_k modelled by replacing

$$\begin{cases} \dots \\ Z_k \quad \leftarrow f_k(\text{parents}(Z_k), \eta_k) \\ \dots \end{cases}$$

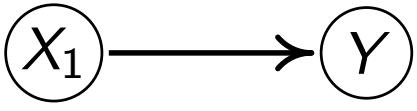
with

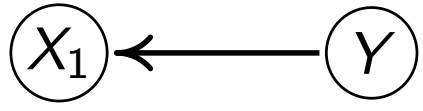
$$\begin{cases} \dots \\ Z_k \quad \leftarrow f_k(\text{parents}(Z_k), \eta_k) + \Delta \\ \dots \end{cases} ,$$

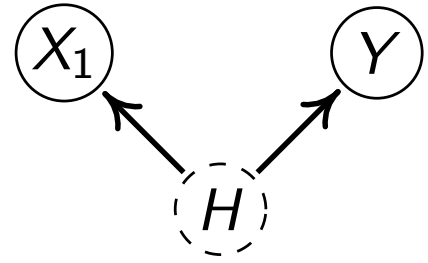
where shift Δ is a deterministic or random.

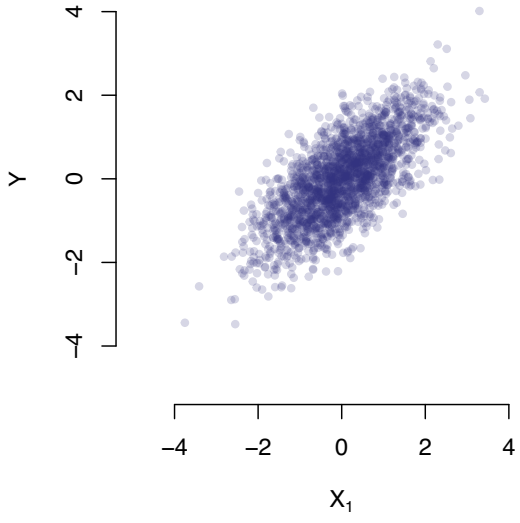
An interesting class of distributions $\mathcal{F}$ consists of all distributions that arise from a SEM under arbitrary interventions on all variables X.

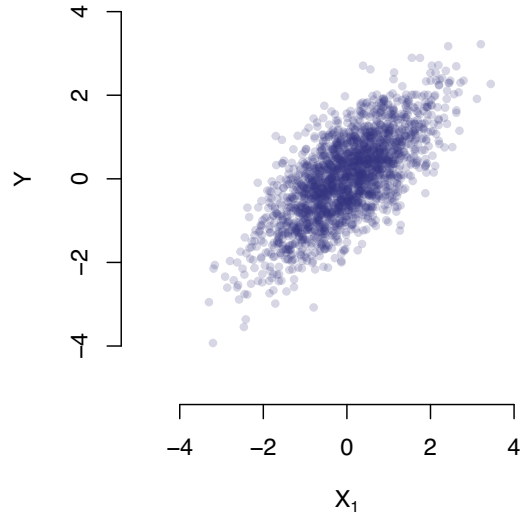
Distribution for observational data

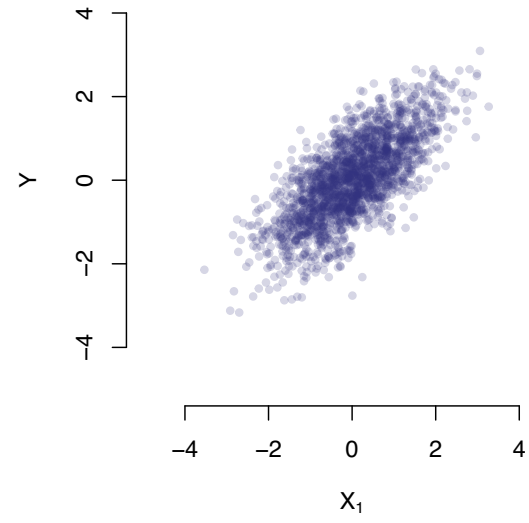




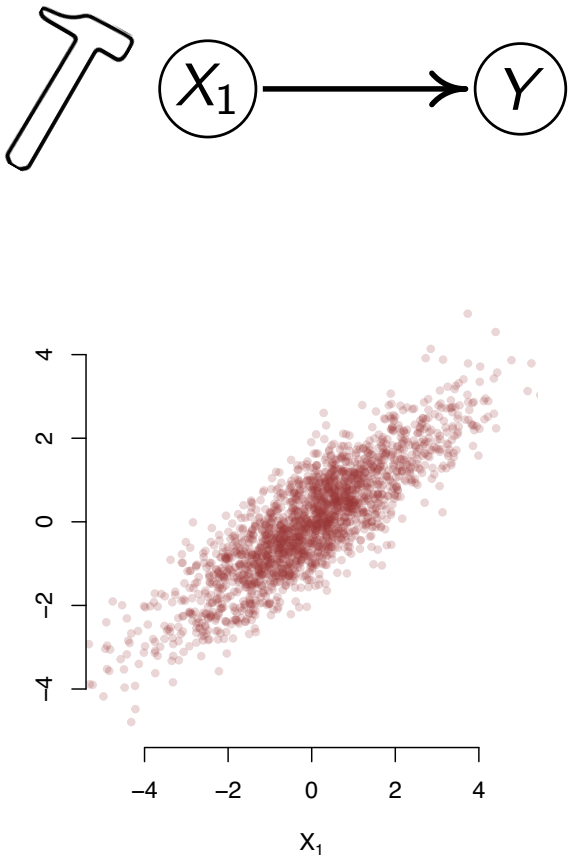


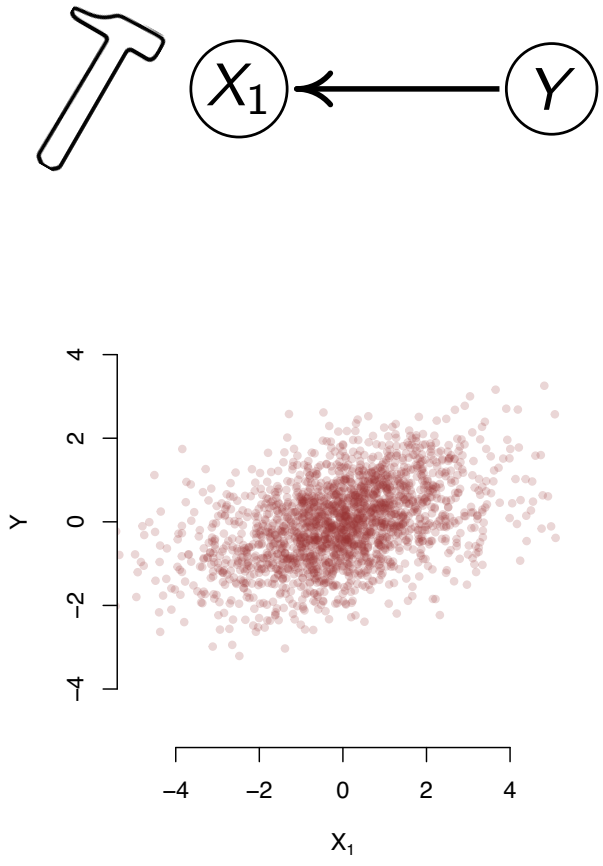


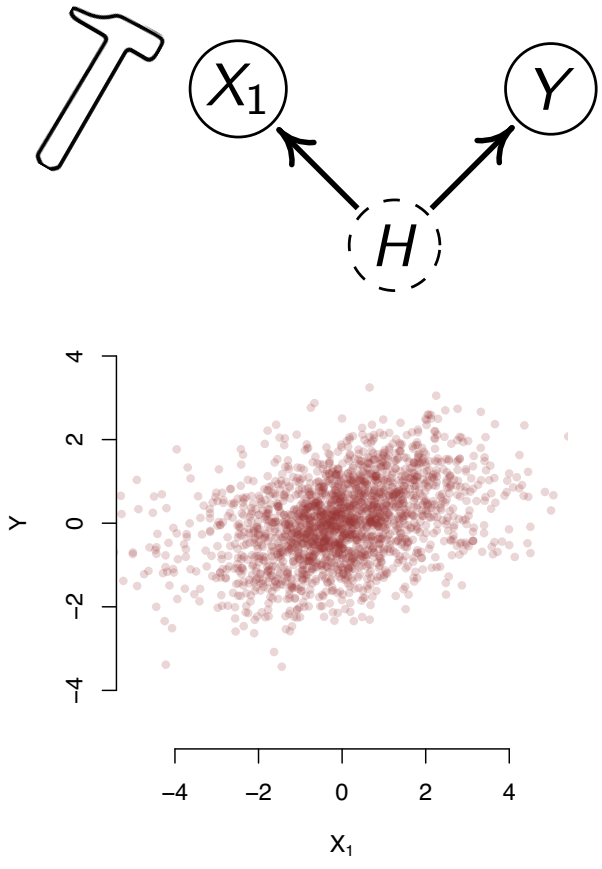




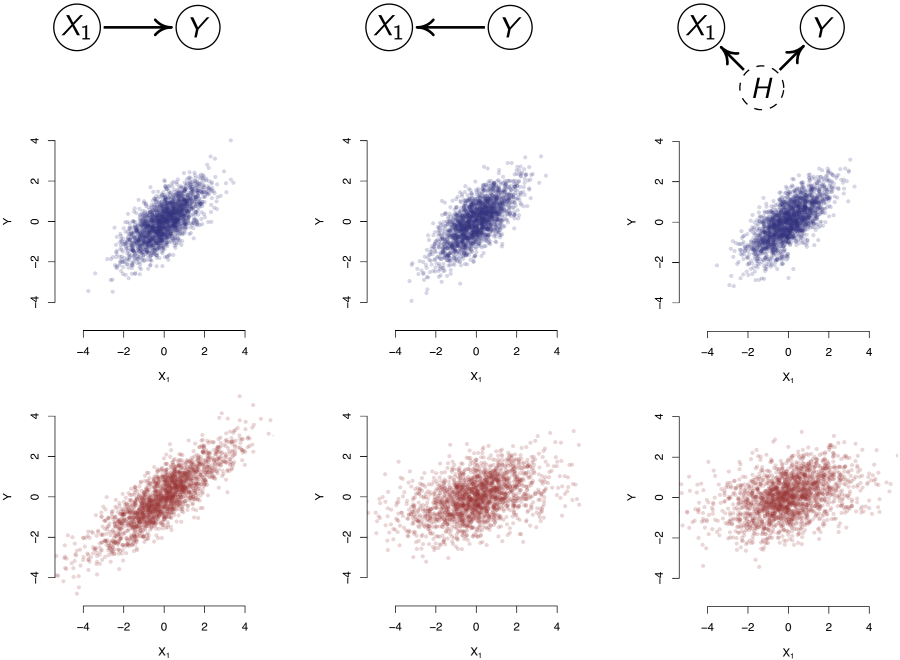
Distribution for shift-intervention on X_1



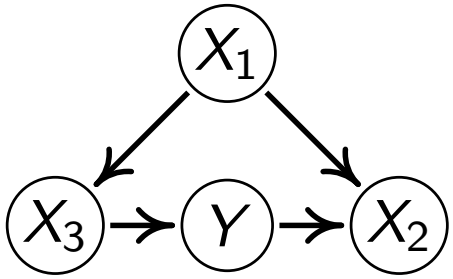




Distribution for observational data and under shift-intervention on X_1



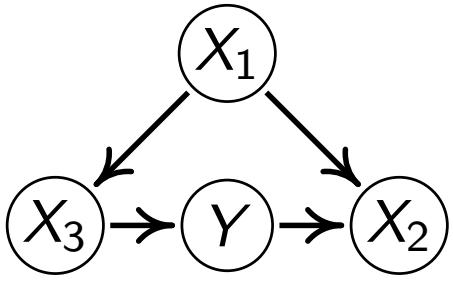
Observational distribution F_{obs}.



$$\begin{cases} X_1 & \leftarrow & & & \eta_1 \\ X_3 & \leftarrow & X_1 & + & \eta_3 \\ Y & \leftarrow & X_3 & + & \eta_y \\ X_2 & \leftarrow & X_1+ & Y+ & \eta_2 \end{cases},$$

and $\eta \sim G$ for some G.

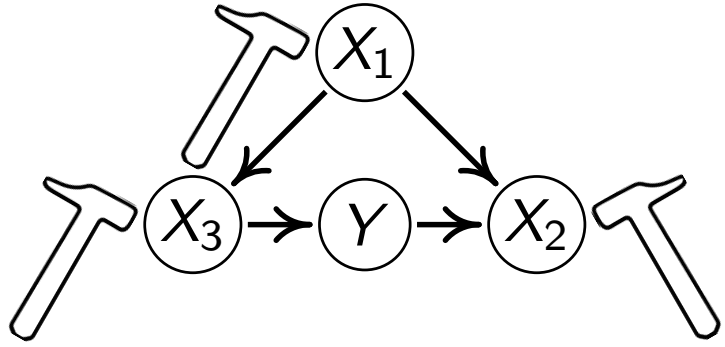
Observational distribution F_{obs}.



$$\begin{cases} X_1 & \leftarrow & & & \eta_1 \\ X_3 & \leftarrow & X_1 & + & \eta_3 \\ Y & \leftarrow & X_3 & + & \eta_y \\ X_2 & \leftarrow & X_1+ & Y+ & \eta_2 \end{cases},$$

and $\eta \sim G$ for some G.

Class $\mathcal{F}$ under interventions

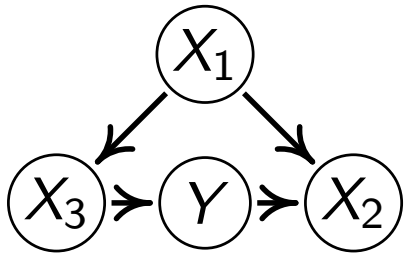


consist eg of all distributions under this SEM with arbitrary shift interventions $\Delta = (\Delta_1, \Delta_2, \Delta_3)$.

$$\begin{cases} X_1 & \leftarrow & & & \eta_1+ & \Delta_1 \\ X_3 & \leftarrow & X_1 & + & \eta_3+ & \Delta_3 \\ Y & \leftarrow & X_3 & + & \eta_y & \\ X_2 & \leftarrow & X_1+ & Y+ & \eta_2+ & \Delta_2 \end{cases},$$

and $\eta \sim G$.

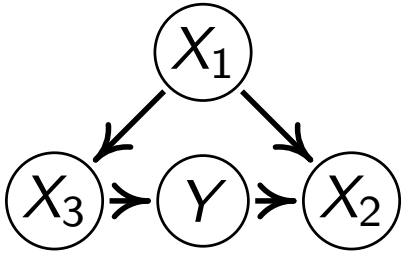
Observational distribution F_{obs}.



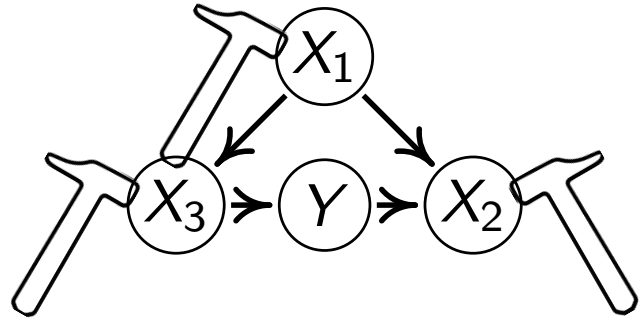
Regression coefficient:

$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} E_F\left[(Y - X'\beta)^2\right]$$
$$= \begin{pmatrix} -1/2 \\ 1/2 \\ 1/2 \end{pmatrix}$$

Observational distribution F_{obs}.



Class $\mathcal{F}$ under interventions



Regression coefficient:

$$\text{argmin}_{\beta \in \mathbb{R}^p} \; E_F\left[(Y - X'\beta)^2\right]$$

$$= \begin{pmatrix} -1/2 \\ 1/2 \\ 1/2 \end{pmatrix}$$

$$\text{argmin}_{\beta \in \mathbb{R}^p} \; \max_{F' \in \mathcal{F}} E_{F'}\left[(Y - X'\beta)^2\right]$$

$$= \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Distribution of residuals $Y - X'\beta$ has to be invariant under a change in distribution of Δ (objective otherwise infinite).

Invariant causal prediction (ICP) of Peters, Bühlmann, M., '16

→ causal effects are regression effects if restricting to the right subset of variables

→ search over all subsets in naive version and keep all for which residual distribution is invariant (null cannot be rejected).

Disadvantages

→ computationally expensive for 'large p'

→ assumes absence of latent confounding

For regression and $F = F_{obs}$,

$$\beta = \text{argmin}_{\beta \in \mathbb{R}^p} \; E_F\left[(Y - X'\beta)^2\right]$$

is equivalent to

$$E_F\Big(\|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty\Big) = 0$$

For regression and $F = F_{obs}$,

$$\beta = \text{argmin}_{\beta \in \mathbb{R}^p} \; E_F\left[(Y - X'\beta)^2\right]$$

is equivalent to

$$E_F\Big(\|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty\Big) = 0$$

For additive interventions

$$\beta = \text{argmin}_{\beta \in \mathbb{R}^p} \; \max_{F' \in \mathcal{F}} E_{F'}\left[(Y - X'\beta)^2\right]$$

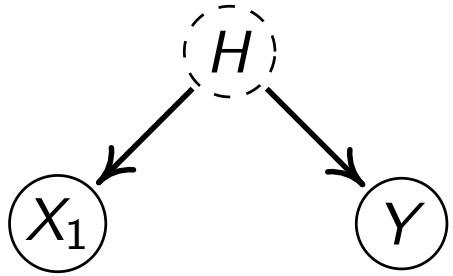
is equivalent to

$$E_{F'}\Big(\|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty\Big) = \text{constant for all } F' \in \mathcal{F}$$

$f(\Delta) + X^0$ η

Latent confounding

Observational distribution

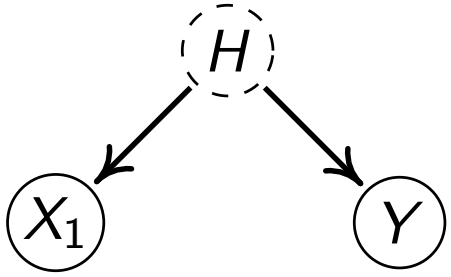


$$\begin{cases} H & \leftarrow & & \eta_h \\ X_1 & \leftarrow & H+ & \eta_1 \\ Y & \leftarrow & H+ & \eta_y \end{cases},$$

and $(\eta_1, \eta_y, \eta_h) \sim \mathcal{N}(0, 1_{3\times 3})$.

Latent confounding

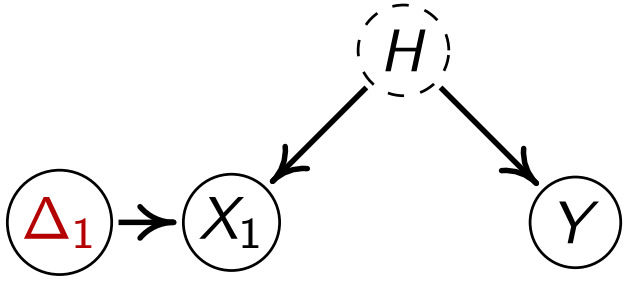
Observational distribution



$$\begin{cases} H & \leftarrow & & \eta_h \\ X_1 & \leftarrow & H+ & \eta_1 \\ Y & \leftarrow & H+ & \eta_y \end{cases},$$

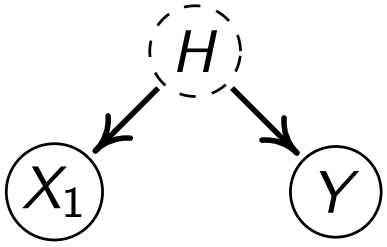
and $(\eta_1, \eta_y, \eta_h) \sim \mathcal{N}(0, 1_{3\times 3})$.

Interventions



Class $\mathcal{F}$ consist of all distributions under this SEM with arbitrary interventions Δ_1 on X_1.

$$\begin{cases} H & \leftarrow & & \eta_h \\ X_1 & \leftarrow & H+ & \eta_1+ \quad \Delta_1 \\ Y & \leftarrow & H+ & \eta_y \end{cases},$$



Optimal regression vector for observational distribution $F = F_{obs}$,

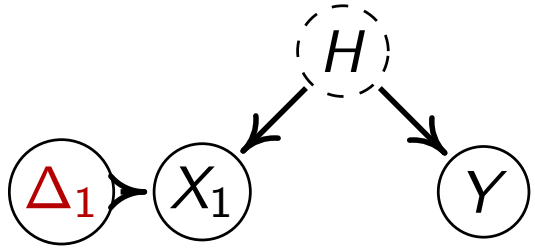
$$\beta = \text{argmin}_{\beta \in \mathbb{R}} \; E_F\left[(Y - X_1\beta)^2\right]$$

$$\Leftrightarrow$$

$$E_F\Big(\|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty\Big) = 0$$

with solution

$$\beta = \frac{1}{2}.$$



Optimal causal coefficient vector for arbitrary distribution F' in class $\mathcal{F}$,

$$\beta = \text{argmin}_{\beta \in \mathbb{R}} \; \max_{F' \in \mathcal{F}} E_{F'}\left[(Y - X_1\beta)^2\right]$$

$$\Leftrightarrow$$

$$E_F\Big(\|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)\|_\infty\Big) \text{ constant}$$

for all $F \in \mathcal{F}$.

with solution

$$\beta = 0.$$

Let $\mathbf{X}_F$ be predictor-matrix under distribution F and same for $\mathbf{Y}_F$. Define

$$\hat{\rho}_F = \mathbf{X}_F' \mathbf{Y}_F \text{ and } \hat{\Sigma}_F = \mathbf{X}_F' \mathbf{X}_F.$$

"Low-dimensional Dantzig" (for observational distribution F):

$$\begin{aligned}
\hat{\beta} &= \text{argmin}_{\beta \in \mathbb{R}^p} \|\mathbf{X}_F'(\mathbf{Y}_F - \mathbf{X}_F \beta)\|_\infty \\
&= \text{argmin}_{\beta \in \mathbb{R}^p} \|\hat{\rho}_F - \hat{\Sigma}_F \beta\|_\infty \\
&= \hat{\Sigma}_F^{-1} \hat{\beta} = \hat{\beta}^{OLS} \text{ (if inverse exists)}
\end{aligned}$$

Let $\mathbf{X}_F$ be predictor-matrix under distribution F and same for $\mathbf{Y}_F$. Define

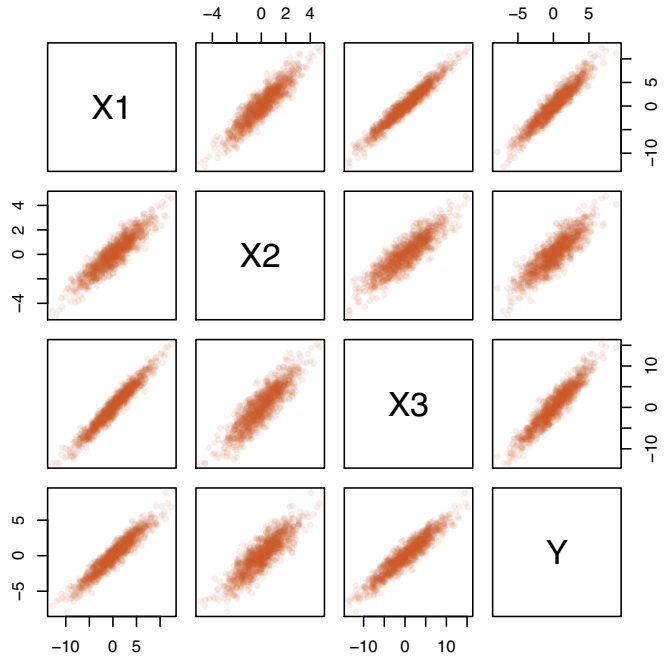
$$\hat{\rho}_F = \mathbf{X}_F' \mathbf{Y}_F \text{ and } \hat{\Sigma}_F = \mathbf{X}_F' \mathbf{X}_F.$$

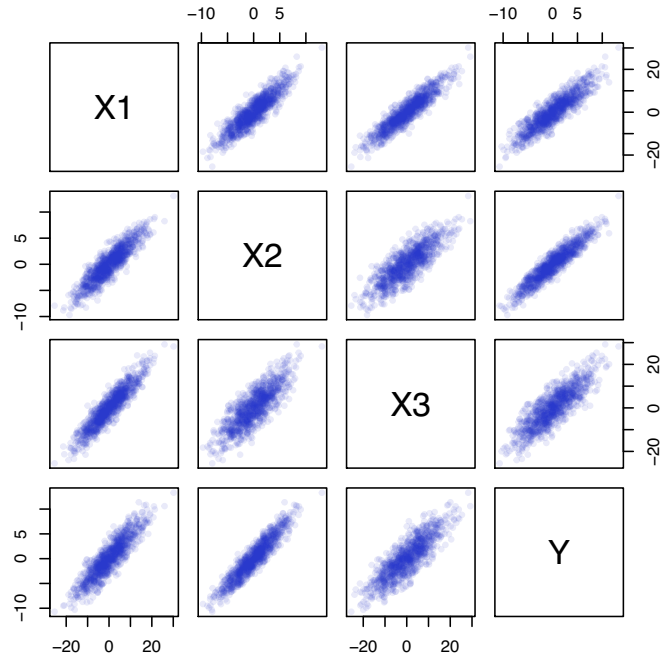
"Low-dimensional Dantzig" (for observational distribution F):

$$\begin{aligned}\hat{\beta} &= \text{argmin}_{\beta \in \mathbb{R}^p} \|\mathbf{X}_F'(\mathbf{Y}_F - \mathbf{X}_F \beta)\|_\infty \\ &= \text{argmin}_{\beta \in \mathbb{R}^p} \|\hat{\rho}_F - \hat{\Sigma}_F \beta\|_\infty \\ &= \hat{\Sigma}_F^{-1} \hat{\beta} = \hat{\beta}^{OLS} \text{ (if inverse exists)}\end{aligned}$$

"Low-dimensional causal Dantzig" (for distributions $F, F' \in \mathcal{F}$):

$$\begin{aligned}\hat{\beta} &= \text{argmin}_{\beta \in \mathbb{R}^p} \| \underbrace{(\hat{\rho}_F - \hat{\rho}_{F'})}_{=:\hat{\delta}} - \underbrace{(\hat{\Sigma}_F - \hat{\Sigma}_{F'})}_{=:\hat{G}} \beta \|_\infty. \\ &= \text{argmin}_{\beta \in \mathbb{R}^p} \|\hat{\delta} - \hat{G}\beta\|_\infty \\ &= \hat{G}^{-1} \hat{\delta} \text{ (if inverse exists)}\end{aligned}$$





Data are generated in two environments $\{1, 2\} = \mathcal{E}$ according to

$$\begin{cases} X_2^e & \leftarrow & & & \eta^0 + & \sigma^e \eta_2 \\ Y^e & \leftarrow & & X_2^e + & \eta^0 + & \eta_y \\ X_1^e & \leftarrow & Y^e + & X_2^e + & & \sigma^e \eta_1 \\ X_3^e & \leftarrow & & X_1^e + & \eta^0 + & \sigma^e \eta_3 \end{cases},$$

where $(\eta^0, \eta_y, \eta_1, \eta_2, \eta_3)$ is assumed to be drawn from $\mathcal{N}_5(0, \mathrm{Id}_5)$ and the noise variances are $\sigma^e = 1$ for environment $e = 1$ and $\sigma^e = 4$ for environment $e = 2$.

Procedure is implemented as method `causalDantzig` in R-package `InvariantCausalPrediction`.

```
> fit <- causalDantzig(X,Y,E,regularization=FALSE)
> print(fit)
Unregularized causal Dantzig
Call:
causalDantzig(X = X, Y = Y, E = E, regularization = FALSE)

   Estimate StdErr p.value
X1   -0.042  0.059   0.481
X2    0.999  0.106  <2e-16 ***
X3    0.035  0.042   0.403
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Low-dimensional causal Dantzig

$\rightarrow$ can be defined for more than two data-environments

$\rightarrow$ can be shown to be asymptotic normal and asymptotically efficient

$\rightarrow$ asymptotic confidence intervals available

$\rightarrow$ works under errors-in-variables as

$$E\big((X+\xi_x)(\eta_y+\xi_y)\big) \;=\; E\big(X\eta_y\big)$$

if η_y is noise of Y and (ξ_x, ξ_y) is additional observation noise in (X, Y).

Comparison to ICP (Peters, Bühlmann, M., '16):

$\rightarrow$ causal Dantzig much more efficient computationally

$\rightarrow$ ICP works for arbitrary interventions (not just shift) but assumes absence of hidden variables

Comparison to ICP (Peters, Bühlmann, M., '16):

$\rightarrow$ causal Dantzig much more efficient computationally

$\rightarrow$ ICP works for arbitrary interventions (not just shift) but assumes absence of hidden variables

Comparison to instrumental variables (IV)

$\rightarrow$ For one-dimensional instrument, IV can estimate only a single causal coefficient ($p = 1$) whereas causal Dantzig can have identifiability for large p (as IV exploits mean shift and causal Dantzig uses second moments)

$\rightarrow$ IV not consistent of the intervention changes the error distribution (for example $X \leftarrow H + (1 + \alpha\Delta)\eta_x$ instead of $X \leftarrow H + \alpha\Delta + \eta_x$).

Binary instrument $e \in \{1, 2\}$. IV in population case can be written as

$$\lim_{n\to\infty} \hat{\beta}_{\mathsf{IV}} = \frac{\mathbb{E}[Y|e=1] - \mathbb{E}[Y|e=2]}{\mathbb{E}[X|e=1] - \mathbb{E}[X|e=2]} = \frac{\mathbb{E}[Y^1] - \mathbb{E}[Y^2]}{\mathbb{E}[X^1] - \mathbb{E}[X^2]}.$$

Causal Dantzig leads to

$$\lim_{n\to\infty} \hat{\beta} = \frac{\mathbb{E}[X^1 \cdot Y^1] - \mathbb{E}[X^2 \cdot Y^2]}{\mathbb{E}[(X^1)^2] - \mathbb{E}[(X^2)^2]}.$$

Consistency	$X = \alpha e + H + \eta_x$ (mean-shift)	$X = H + (1 + \alpha e)\eta_x$ (change in error distribution)
Instrumental variable regression	yes	no
Unregularized causal Dantzig	yes	yes

"High-dimensional Dantzig" (Candes and Tao, '07)

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \|\hat{\rho}_F - \hat{\Sigma}_F \beta\|_\infty \leq \lambda$$

"High-dimensional Dantzig" (Candes and Tao, '07)

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \|\hat{\rho}_F - \hat{\Sigma}_F \beta\|_\infty \leq \lambda$$

"High-dimensional (large p) causal Dantzig"

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \|\hat{\delta} - \hat{G} \beta\|_\infty \leq \lambda$$

"High-dimensional Dantzig" (Candes and Tao, '07)

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \|\hat{\rho}_F - \hat{\Sigma}_F \beta\|_\infty \leq \lambda$$

"High-dimensional (large p) causal Dantzig"

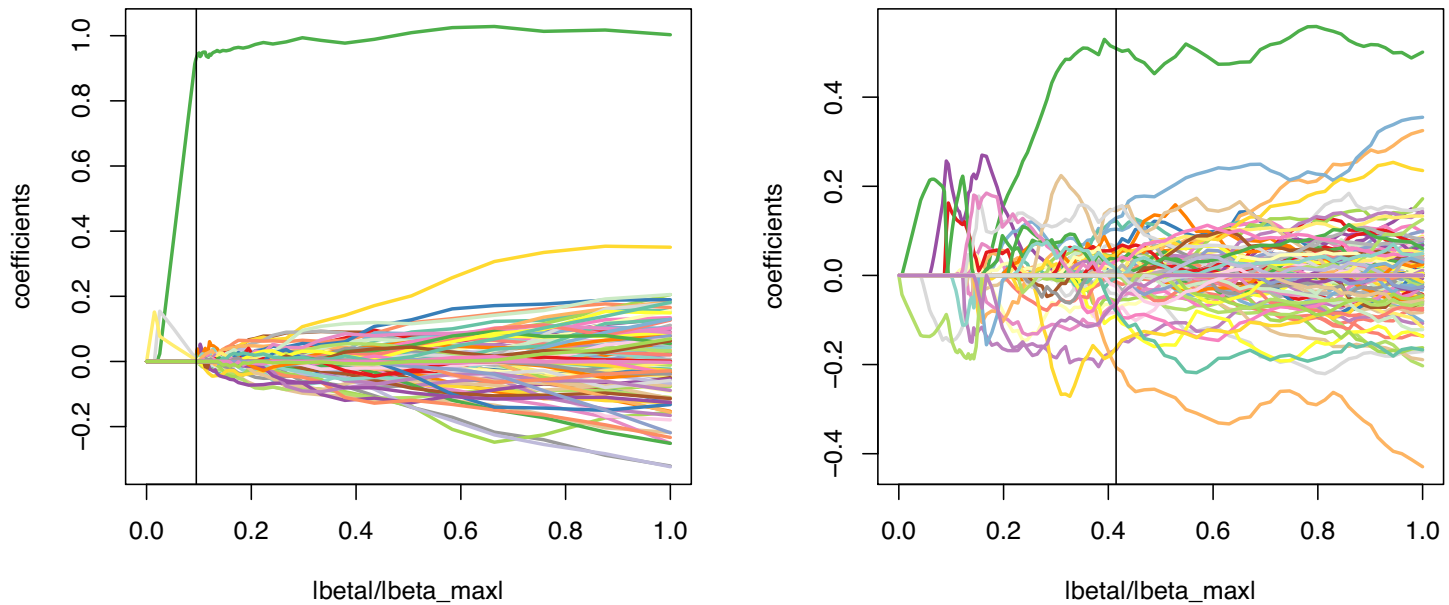
$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \|\hat{\delta} - \hat{G} \beta\|_\infty \leq \lambda$$

$\rightarrow$ penalty can be chosen by cross-validation with ℓ_∞-objective

$\rightarrow$ for more than two distributions

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \|\beta\|_1 \text{ such that } \max_{F, F' \in \mathcal{F}} \|(\hat{\rho}_F - \hat{\rho}_{F'}) - (\hat{\Sigma}_F - \hat{\Sigma}_{F'})\beta\|_\infty \leq \lambda$$

Regularization paths
(true $\beta = (1, 0, 0, 0, \ldots)$, first component green)



Left: $p = 100$, $n = 200$. Right: $p = 200$, $n = 60$.

Ye and Zhang ('10) analysis of Dantzig selector.

Cone invertibility factor is a lower bound on the ℓ_∞-norm of $\hat{\mathbf{\Sigma}}u$, given that u lies in the cone $\{u : \|u_{S^c}\|_1 \leq \|u_S\|_1\}$ and has unit norm $\|u\|_q = 1$.

$$\mathrm{CIF}_q(S) = \inf_u \left\{ \frac{|S|^{1/q}\|\hat{\mathbf{\Sigma}}u\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \leq \|u_S\|_1 \right\}.$$

Ye and Zhang ('10) analysis of Dantzig selector.

Cone invertibility factor is a lower bound on the ℓ_∞-norm of $\hat{\mathbf{\Sigma}}u$, given that u lies in the cone $\{u : \|u_{S^c}\|_1 \leq \|u_S\|_1\}$ and has unit norm $\|u\|_q = 1$.

$$\mathrm{CIF}_q(S) = \inf_u \left\{ \frac{|S|^{1/q}\|\hat{\mathbf{\Sigma}}u\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \leq \|u_S\|_1 \right\}.$$

Define *causal cone invertibility factor* $\mathrm{CCIF}_q(S, \hat{\mathbf{G}})$ as

$$\mathrm{CCIF}_q(S, \hat{\mathbf{G}}) := \inf_u \left\{ \frac{|S|^{1/q}\|\hat{\mathbf{G}}u\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \leq \|u_S\|_1 \right\}.$$

Assumptions (for two distributions indexed by $e \in \{1, 2\}$)

(i) inner-product invariance holds for $(X^e, Y^e), e \in \{1, 2\}$ under a β^*

(ii) $X^1, X^2, \eta^1_{p+1}, \eta^2_{p+1}$ are centered and multivariate Gaussian.

(iii) all involved error variances are bounded by σ^2.

Let $\lambda \asymp 5C\sqrt{\log(p)/\min_{e\in\{1,2\}} n_e}$ for a constant $C > 0$ that satisfies $\sigma < C < \infty$. With probability converging to 1 as $n_1, n_2, p \to \infty$,

$$\|\hat{\beta}^\lambda - \beta^*\|_q \leq \frac{10C}{\mathrm{CCIF}_q(S, \hat{\mathbf{G}})} |S|^{1/q} \sqrt{\frac{\log(p)}{\min_{e\in\{1,2\}} n_e}}$$

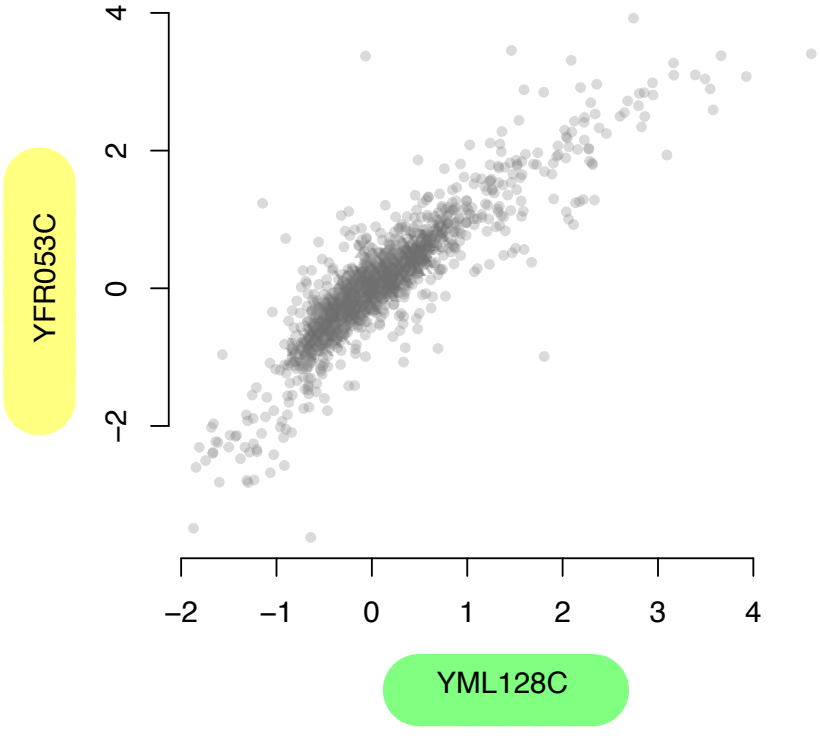
Let $\hat{S}$ denote the active set of $\hat{\beta}^{\lambda}$. Assume a betamin-type condition

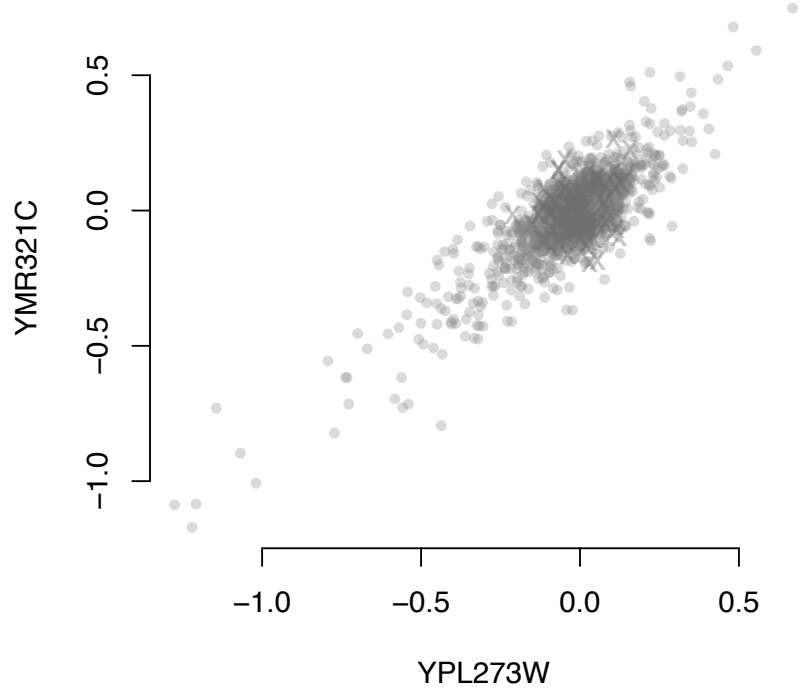
$$\min_{k \in S} |\beta_k^*| > \frac{10C}{\mathrm{CCIF}_{\infty}(S, \hat{\mathbf{G}})} \sqrt{\frac{\log(p)}{\min_{e \in \{1,2\}} n_e}}.$$

Then under the previous assumptions for $q = \infty$, we have

$$\mathbb{P}[\hat{S} \supseteq S] \to 1 \qquad \text{for } n_1, n_2, p \to \infty.$$

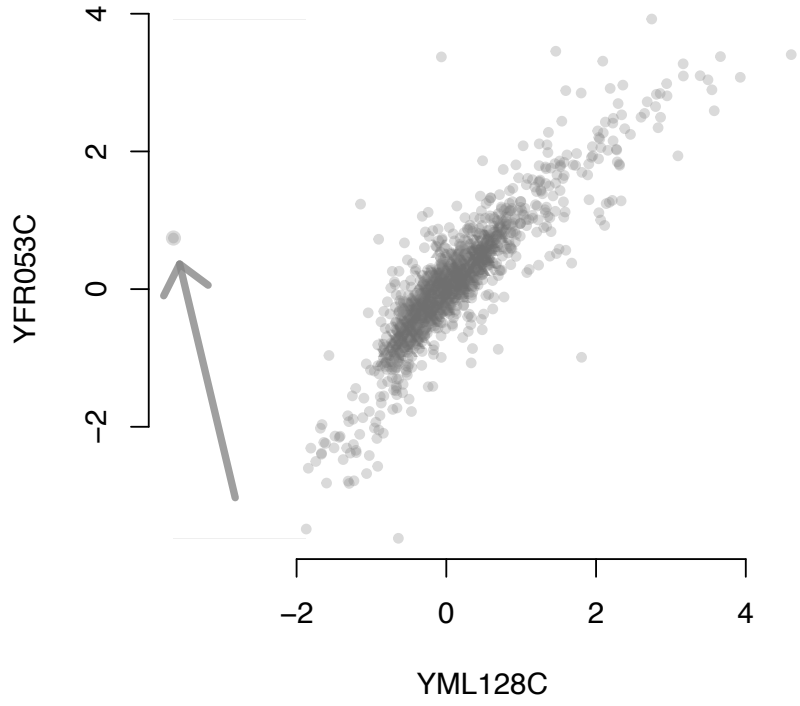
Kemmeren data: gene activities for around 5000 genes in yeast.

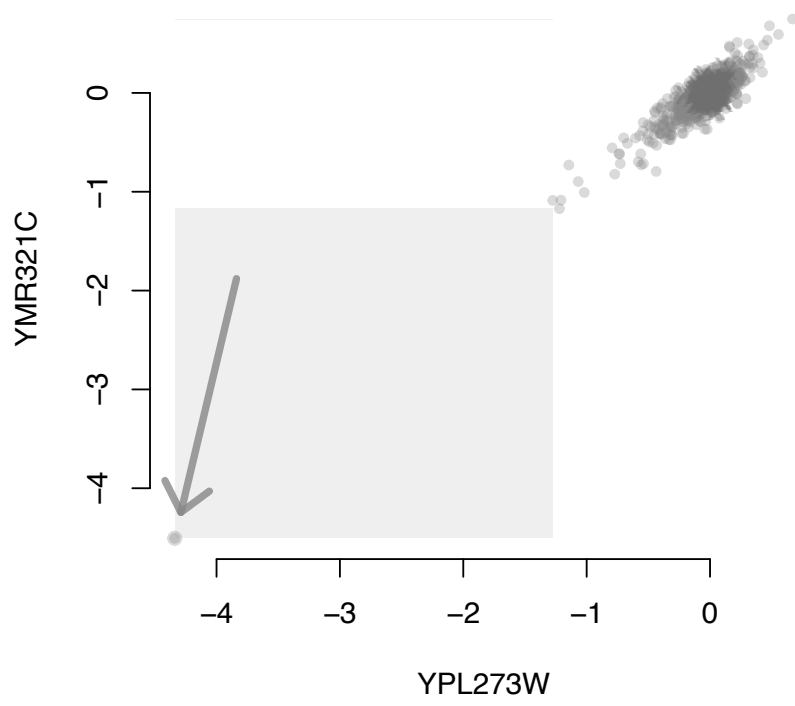




What happens to YFR053C if we knock out/delete gene YML128C?
What happens to YMR321C if we knock out/delete gene YPL273W?

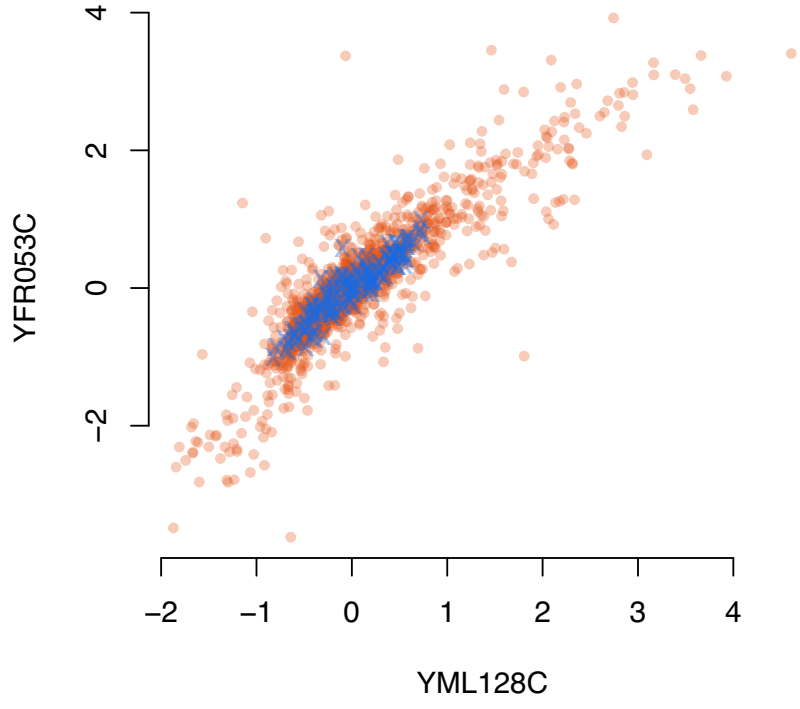
We can check by looking at interventions.

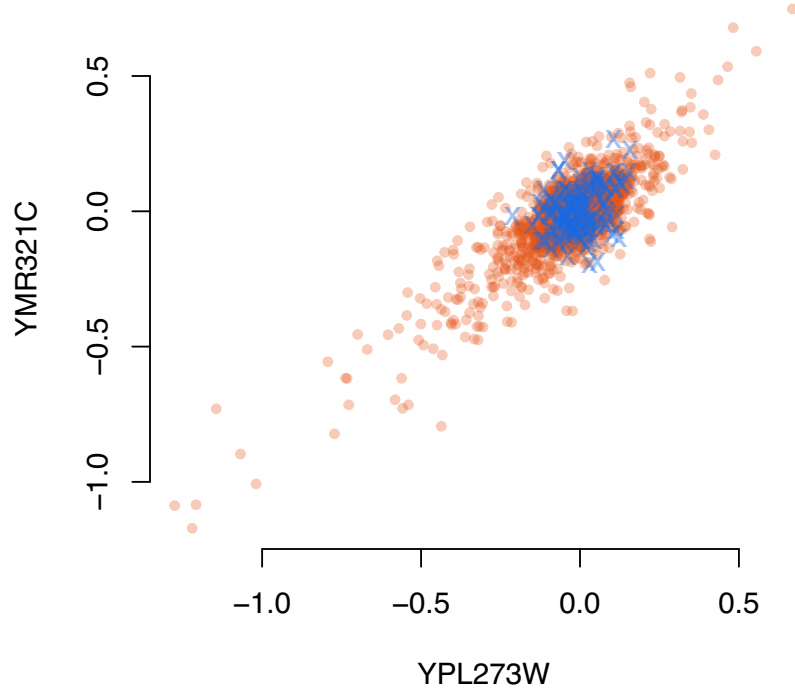




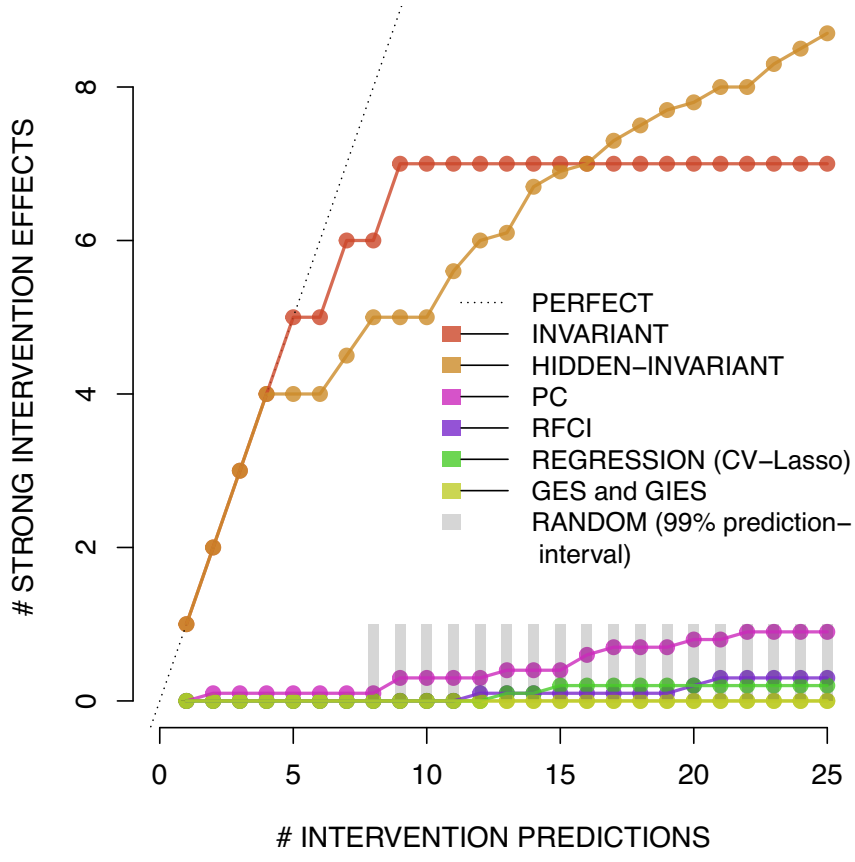
But could we have predicted these outcomes just from the previous data?

divide into observational ($n_{obs} = 160$) and interventional data ($n_{int} \approx 1600$)





What happens to YFR053C if we knock out/delete gene YML128C?
What happens to YMR321C if we knock out/delete gene YPL273W?


PERFECT
INVARIANT
HIDDEN−INVARIANT
PC
RFCI
REGRESSION (CV−Lasso)
GES and GIES
RANDOM (99% prediction−
interval)
STRONG INTERVENTION EFFECTS
0
2
4
6
8
0
5
10
15
20
25
INTERVENTION PREDICTIONS

E. Candes and T. Tao (2007)
The Dantzig selector: statistical estimation when p is much larger than n.
Annals of Statistics, 35:2313–2351

F. Ye and C.-H. Zhang (2010)
Rate minimaxity of the Lasso and Dantzig selector for the lq loss in lr balls.
Journal of Machine Learning Research, 11:3519– 3540

J. Peters, P. Bühlmann, and M. (2016)
Causal inference by using invariant prediction: identification and confidence intervals.
Journal of the Royal Statistical Society, Series B (with discussion) 78:947–1012

M., A. Hauser, J.M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann (2016)
Methods for causal inference from gene perturbation experiments and validation.
Proceedings of the National Academy of Sciences 113: 7361–7368

D. Rothenhäusler, P. Bühlmann, M. (2017)
Causal Dantzig: fast inference in linear structural equation models with hidden variables under additive interventions
arxiv.org/1706.06159

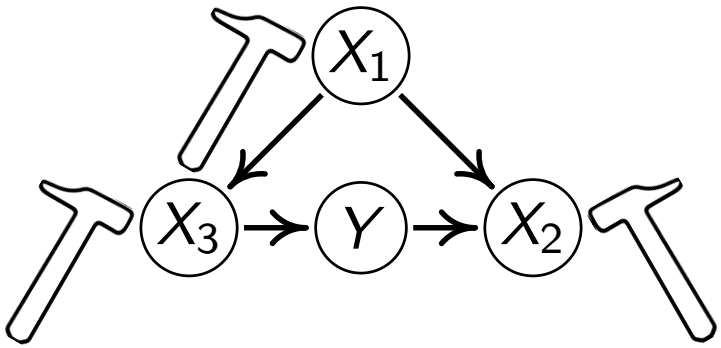
R-package `InvariantCausalPrediction` on CRAN.

Thank you!

Let $S^* \subseteq \{1, \ldots, p\}$ be causal parents of Y:

$$S^* = \{k : \beta_k^* \neq 0\}.$$

In absence of hidden confounding, causal vector β^* is optimal regression vector when regressing Y on X_{S^*}.



The residual distribution of $Y - X\beta^*$ is invariant across all $F, F' \in \mathcal{F}$.

...exploit this fact to estimate S^* and β^*.

In example: $Y - X_3\beta_3^*$ is invariant under a change in Δ.

Sketch of ICP, invariant causal prediction (Peters, Bühlmann, M. '16):

For all $S \subseteq \{1, \ldots, p\}$,

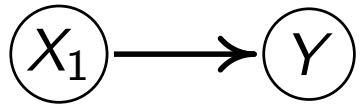
(i) Estimate optimal regression $\hat{\beta}^{(S)}$ vector for $Y \sim X_S$.

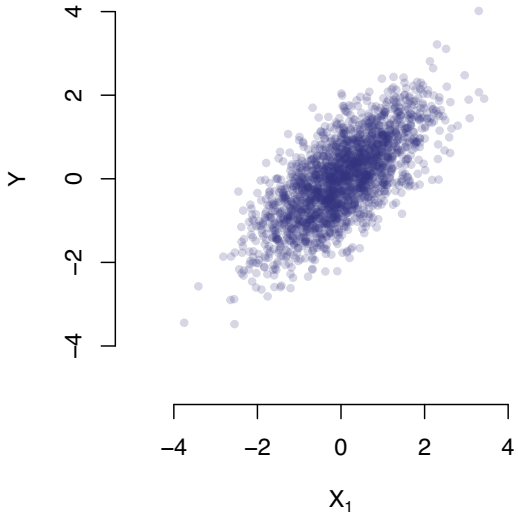
(ii) Test whether residual distribution $Y - X_S \beta^{(S)}$ is identical for different distributions/data sets $F, F' \in \mathcal{F}$.

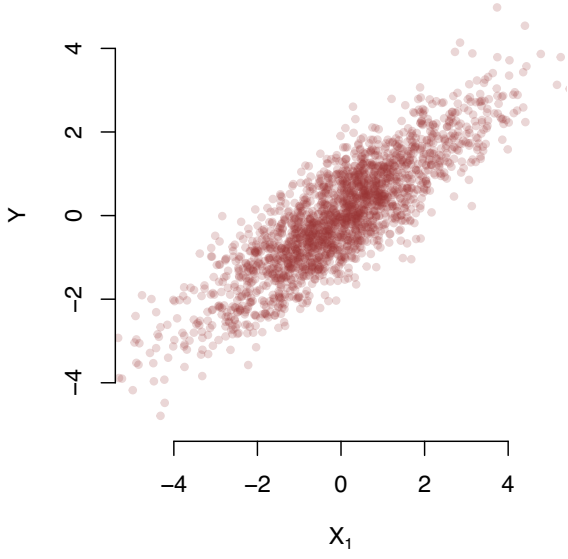
Define

$$\hat{S} = \bigcap_{S\text{: null accepted}} S.$$

FWER-control follows: $P(\hat{S} \subseteq S^*) \geq 1 - \alpha$.

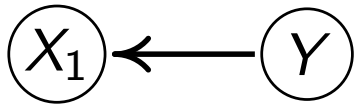


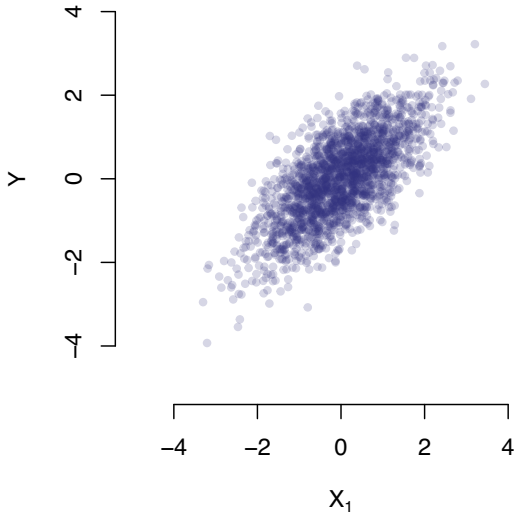


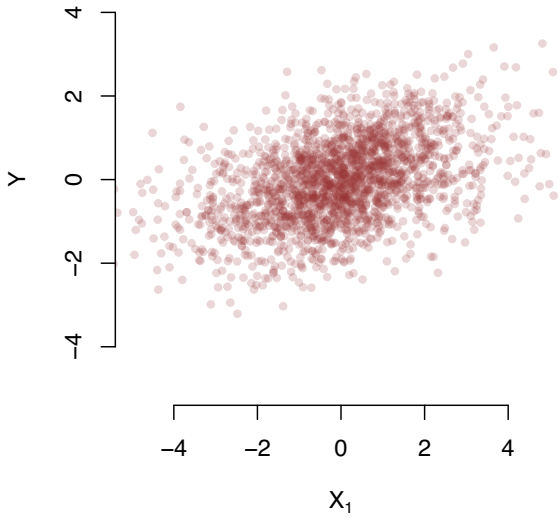


S	$H_{0,S}$ true ?
$S = \emptyset$	no
$S = \{X_1\}$	yes

$$\bigcap_{S:H_{0,S} \text{ true}} S = \{X_1\}.$$







S	$H_{0,S}$ true ?
$S = \emptyset$	yes
$S = \{X_1\}$	no

$$\bigcap_{S:H_{0,S} \text{ true}} S = \emptyset.$$