Learning Optimal Interventions

Jonas Mueller*, David Reshef, George Du, Tommi Jaakkola

*jonasmueller@csail.mit.edu

MIT Computer Science & Artificial Intelligence Laboratory

Dec 10, 2016

Problem Setup

Goal: Identify beneficial *interventions* from limited (observational) data

Problem Setup

Goal: Identify beneficial *interventions* from limited (observational) data

- Dataset $\mathcal{D}_n := \left\{ \left(x^{(i)}, y^{(i)} \right) \right\}_{i=1}^{n} \overset{iid}{\sim} P_{XY}$

  $X \in \mathbb{R}^d =$ covariates (features) of individual

  $Y \in \mathbb{R} =$ outcome of interest

Problem Setup

Goal: Identify beneficial *interventions* from limited (observational) data

- Dataset $\mathcal{D}_n := \left\{ \left(x^{(i)}, y^{(i)} \right) \right\}_{i=1}^{n} \overset{iid}{\sim} P_{XY}$

  $X \in \mathbb{R}^d =$ covariates (features) of individual

  $Y \in \mathbb{R} =$ outcome of interest

- **Objective:** Influence X to produce (expected) improvement in Y
  (requires simplifying causal assumptions)

Problem Setup

Goal: Identify beneficial *interventions* from limited (observational) data

- Dataset $\mathcal{D}_n := \left\{ \left(x^{(i)}, y^{(i)} \right) \right\}_{i=1}^{n} \overset{iid}{\sim} P_{XY}$

  $X \in \mathbb{R}^d =$ covariates (features) of individual

  $Y \in \mathbb{R} =$ outcome of interest

- **Objective:** Influence X to produce (expected) improvement in Y

  (requires simplifying causal assumptions)

- Among feasible transformations to X, which one is best?

  Limited data $\implies$ inherent uncertainty regarding $Y \mid X$ relationship

Assumptions

(A1) Underlying graphical model: $X \rightarrow \widetilde{X} \rightarrow Y$

$X \sim P_X =$ pre-intervention covariate-values

$\widetilde{X} =$ values after performing a chosen intervention

Assumptions

(A1) Underlying graphical model: $X \to \widetilde{X} \to Y$

$X \sim P_X$ = pre-intervention covariate-values

$\tilde{X}$ = values after performing a chosen intervention

(A2) Under no intervention: $\widetilde{X} = X$ (and in data $\mathcal{D}_n$: $\widetilde{x}_i = x_i$)

Assumptions

(A1) Underlying graphical model: $X \to \widetilde{X} \to Y$

$X \sim P_X$ = pre-intervention covariate-values

$\widetilde{X}$ = values after performing a chosen intervention

(A2) Under no intervention: $\widetilde{X} = X$ (and in data $\mathcal{D}_n$: $\widetilde{x}_i = x_i$)

(A3) $\widetilde{X} = T(X)$ (Intervention can be precisely enacted)

$T : \mathbb{R}^d \to \mathbb{R}^d$ = desired transformation of covariate-values (to guide intervention)

Assumptions

(A1) Underlying graphical model: $X \to \widetilde{X} \to Y$

$X \sim P_X$ = pre-intervention covariate-values

$\tilde{X}$ = values after performing a chosen intervention

(A2) Under no intervention: $\widetilde{X} = X$ (and in data $\mathcal{D}_n$: $\widetilde{x}_i = x_i$)

(A3) $\widetilde{X} = T(X)$ (Intervention can be precisely enacted)

$T : \mathbb{R}^d \to \mathbb{R}^d$ = desired transformation of covariate-values (to guide intervention)

(A4) $Y = f(\widetilde{X}) + \varepsilon$ (with $\mathbb{E}[\varepsilon] = 0, \varepsilon \perp\!\!\!\perp \widetilde{X}, X$)

Invariant relationship[1]: Same f for $\widetilde{X}$ produced by any (or no) intervention

[1] Peters J, Bühlmann P, Meinshausen N. Causal inference using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B* (2016)

Overview of Framework

Identifying intervention $=$ find desired transformation policy T

- $\widetilde{x} = T(x) \in \mathcal{C}_x$: post-intervention covariate-measurements of individual with initial measurements $x \in \mathbb{R}^d$, for intervention to enact T, $f(T(x)) = \mathbb{E}_\varepsilon[Y \mid \widetilde{X} = T(x)]$
- $\mathcal{C}_x \subset \mathbb{R}^d$: constraints on possible transformations of x
- $\mathcal{C}_x := \{z \in \mathbb{R}^d : |x_s - z_s| \leqslant \gamma_s\} \implies s^{\text{th}}$ feature cannot be altered by more than γ_s
- $\mathcal{C}_x := \{z \in \mathbb{R}^d : ||x - z||_0 \leqslant k\} \implies$ at most k features can be intervened upon

Overview of Framework

Identifying intervention = find desired transformation policy T

(Step 1) Bayesian inference of posterior $f \mid \mathcal{D}_n$ (eg. Gaussian Process)

Summarized by mean, covariance functions: $\mathbb{E}[f(x) \mid \mathcal{D}_n]$, Cov$([f(x)\ f(x')] \mid \mathcal{D}_n)$

- $\widetilde{x} = T(x) \in \mathcal{C}_x$: post-intervention covariate-measurements of individual with initial measurements $x \in \mathbb{R}^d$, for intervention to enact T, $f(T(x)) = \mathbb{E}_\varepsilon[Y \mid \widetilde{X} = T(x)]$
- $\mathcal{C}_x \subset \mathbb{R}^d$: constraints on possible transformations of x
- $\mathcal{C}_x := \{z \in \mathbb{R}^d : |x_s - z_s| \leqslant \gamma_s\} \implies s^{\text{th}}$ feature cannot be altered by more than γ_s
- $\mathcal{C}_x := \{z \in \mathbb{R}^d : ||x - z||_0 \leqslant k\} \implies$ at most k features can be intervened upon

Overview of Framework

Identifying intervention $=$ find desired transformation policy T

(Step 1) Bayesian inference of posterior $f \mid \mathcal{D}_n$ (eg. Gaussian Process)

Summarized by mean, covariance functions: $\mathbb{E}[f(x) \mid \mathcal{D}_n]$, $\text{Cov}([f(x)\ f(x')] \mid \mathcal{D}_n)$

(Step 2) Optimize of T w.r.t. posterior $f \mid \mathcal{D}_n$ (subject to $T(x) \in \mathcal{C}_x$)

to identify feasible covariate-transformation likely to improve expected outcomes $(f(T(x)) > f(x))$ according to our current beliefs given limited data

- $\widetilde{x} = T(x) \in \mathcal{C}_x$: post-intervention covariate-measurements of individual with initial measurements $x \in \mathbb{R}^d$, for intervention to enact T, $f(T(x)) = \mathbb{E}_\varepsilon[Y \mid \widetilde{X} = T(x)]$
- $\mathcal{C}_x \subset \mathbb{R}^d$: constraints on possible transformations of x
- $\mathcal{C}_x := \{z \in \mathbb{R}^d : |x_s - z_s| \leqslant \gamma_s\} \implies s^{\text{th}}$ feature cannot be altered by more than γ_s
- $\mathcal{C}_x := \{z \in \mathbb{R}^d : ||x - z||_0 \leqslant k\} \implies$ at most k features can be intervened upon

Personalized Intervention

Given new individual with covariate-values $x \in \mathbb{R}^d$, $T(x)$ personally tailored to best improve this individual's expected post-intervention outcome

Personalized Intervention

Given new individual with covariate-values $x \in \mathbb{R}^d$, $T(x)$ personally tailored to best improve this individual's expected post-intervention outcome

Expected individual gain: $\quad G_x(T) := f(T(x)) - f(x) \mid \mathcal{D}_n$

Optimal personalized intervention

Given by optimization of $T(x) \in \mathbb{R}^d$*:* $T^*(x) = \underset{T(x) \in \mathcal{C}_x}{\operatorname{argmax}} \; F^{-1}_{G_x(T)}(\alpha)$

- $F^{-1}_{G(\cdot)}(\alpha)$= α^{th} quantile of posterior distribution for gain function

Personalized Intervention

Given new individual with covariate-values $x \in \mathbb{R}^d$, $T(x)$ personally tailored to best improve this individual's expected post-intervention outcome

Expected individual gain: $\quad G_x(T) := f(T(x)) - f(x) \mid \mathcal{D}_n$

Optimal personalized intervention

Given by optimization of $T(x) \in \mathbb{R}^d$*:* $\;T^*(x) = \underset{T(x) \in \mathcal{C}_x}{\operatorname{argmax}} \; F^{-1}_{G_x(T)}(\alpha)$

- $F^{-1}_{G(\cdot)}(\alpha)$= α^{th} quantile of posterior distribution for gain function
- Posterior for $G_x(T)$ summarized by
  mean $= \mathbb{E}[f(T(x) \mid \mathcal{D}_n] - \mathbb{E}[f(x) \mid \mathcal{D}_n]$
  variance $= \mathsf{Var}(f(T(x)) \mid \mathcal{D}_n) + \mathsf{Var}(f(x) \mid \mathcal{D}_n) - \underbrace{2\mathsf{Cov}(f(T(x)), f(x) \mid \mathcal{D}_n)}_{\text{ties uncertainty at } x \text{ and } T(x)}$

Optimal Personalized Intervention

- $T^*(x)$ improves expected outcome with probability $\geqslant 1 - \alpha$ under our posterior beliefs (conservatively choose $\alpha < 0.5$)

Optimal Personalized Intervention

- $T^*(x)$ improves expected outcome with probability $\geqslant 1 - \alpha$ under our posterior beliefs (conservatively choose $\alpha < 0.5$)

- Will never consider T where $\mathbb{E}[f(T(x) \mid \mathcal{D}_n] < \mathbb{E}[f(x) \mid \mathcal{D}_n]$

  Feasible choice $T(x) = x$ produces objective value of 0

Optimal Personalized Intervention

- $T^*(x)$ improves expected outcome with probability $\geqslant 1 - \alpha$ under our posterior beliefs (conservatively choose $\alpha < 0.5$)

- Will never consider T where $\mathbb{E}[f(T(x) \mid \mathcal{D}_n] < \mathbb{E}[f(x) \mid \mathcal{D}_n]$

  Feasible choice $T(x) = x$ produces objective value of 0

- If α is small & uncertainty is high at x (outlier), then $T^*(x) = x$

  Philosophy: Doing nothing is greatly preferred to causing harm.

  Only propose interventions we are certain will lead to improvement

Intervening on Populations

- Single transformation-policy to improve outcomes for new (or all) individuals sampled from same population as $\mathcal{D}_n$

Intervening on Populations

- Single transformation-policy to improve outcomes for new (or all) individuals sampled from same population as $\mathcal{D}_n$
- May no longer measure features of new individuals

Intervening on Populations

- Single transformation-policy to improve outcomes for new (or all) individuals sampled from same population as $\mathcal{D}_n$
- May no longer measure features of new individuals

Expected population gain: $G_X(T) := \mathbb{E}_X[G_x(T)]$

Empirical estimate: $G_n(T) := \frac{1}{n}\sum_{i=1}^{n}\left[f(T(x^{(i)})) - f(x^{(i)})\right] \ \mid \mathcal{D}_n]$

Intervening on Populations

- Single transformation-policy to improve outcomes for new (or all) individuals sampled from same population as $\mathcal{D}_n$
- May no longer measure features of new individuals

Expected population gain: $G_X(T) := \mathbb{E}_X[G_x(T)]$

Empirical estimate: $G_n(T) := \frac{1}{n} \sum_{i=1}^{n} \left[f(T(x^{(i)})) - f(x^{(i)}) \right] \mid \mathcal{D}_n]$

Optimal population intervention

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmax}} \; F^{-1}_{G_X(T)}(\alpha)$$

- $\mathcal{T} := \{T : \; T(x) \in \mathcal{C}_x \; \forall x\}$ (set of feasible policies)

Intervening on Populations

- Single transformation-policy to improve outcomes for new (or all) individuals sampled from same population as $\mathcal{D}_n$
- May no longer measure features of new individuals

Expected population gain: $G_X(T) := \mathbb{E}_X[G_x(T)]$

Empirical estimate: $G_n(T) := \frac{1}{n}\sum_{i=1}^{n}\left[f(T(x^{(i)})) - f(x^{(i)})\right] \mid \mathcal{D}_n]$

Optimal population intervention

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmax}} \; F^{-1}_{G_X(T)}(\alpha)$$

- $\mathcal{T} := \{T : \; T(x) \in \mathcal{C}_x \; \forall x\}$ (set of feasible policies)

Posterior for $G_n(T)$ has: mean $= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(T(x^{(i)})) \mid \mathcal{D}_n] - \mathbb{E}[f(x^{(i)}) \mid \mathcal{D}_n]$

$$\text{variance} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\text{Cov}\left(f(x^{(i)}), f(x^{(j)}) \mid \mathcal{D}_n\right) - \text{Cov}\left(f(T(x^{(i)})), f(x^{(j)}) \mid \mathcal{D}_n\right) - \text{Cov}\left(f(x^{(i)}), f(T(x^{(j)})) \mid \mathcal{D}_n\right) + \text{Cov}\left(f(T(x^{(i)})), f(T(x^{(j)})) \mid \mathcal{D}_n\right)\Big]$$

Types of Global Policy

- Form of T cannot depend on x

Types of Global Policy

- Form of T cannot depend on x
- *Sparse* intervention: Assume only covariates in chosen intervention-subset $\mathcal{I} \subset \{1, \ldots, d\}$ are changed
  (all other covariates remain fixed at their pre-intervention values)

Types of Global Policy

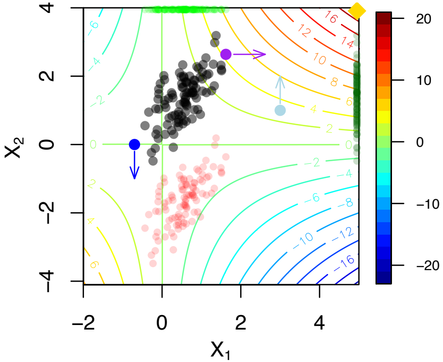
- Form of T cannot depend on x
- *Sparse* intervention: Assume only covariates in chosen intervention-subset $\mathcal{I} \subset \{1, \ldots, d\}$ are changed
  (all other covariates remain fixed at their pre-intervention values)
- *Shift* intervention: $T(x) = x + \Delta$
  $\Delta \in \mathbb{R}^d$ = shift that the policy applies to each individuals' features
  (eg. $T(x) = [x_1, x_2 + 3, \ldots, x_d]$)

Types of Global Policy

- Form of T cannot depend on x
- *Sparse* intervention: Assume only covariates in chosen intervention-subset $\mathcal{I} \subset \{1, \ldots, d\}$ are changed
  (all other covariates remain fixed at their pre-intervention values)
- *Shift* intervention: $T(x) = x + \Delta$
  $\Delta \in \mathbb{R}^d =$ shift that the policy applies to each individuals' features
  (eg. $T(x) = [x_1, x_2 + 3, \ldots, x_d]$)
- *Uniform* intervention: $T(x) = [z_1, \ldots, z_d]$ where $z_j = x_j \ \forall j \notin \mathcal{I}$
  Sets certain covariates to the same constant value for all individuals
  (eg. $T(x) = [x_1, 0, x_3, \ldots, x_d]$)

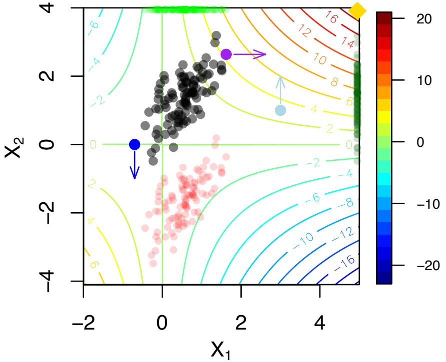
Example: Different Types of Intervention

Contours of outcomes Y expected across feature space $[X_1, X_2]$ if $f(X) = X_1 \cdot X_2$



Example: Different Types of Intervention

Contours of outcomes Y expected across feature space $[X_1, X_2]$ if $f(X) = X_1 \cdot X_2$



- Under sparsity constraint, we must carefully model the underlying population in order to identify best uniform intervention

Algorithms

- Standard GP prior for $f \implies F_{G(T)}^{-1}(\alpha)$ has closed form
- Smooth kernel $\implies$ our objectives differentiable w.r.t. T

Algorithms

- Standard GP prior for $f \implies F^{-1}_{G(T)}(\alpha)$ has closed form
- Smooth kernel $\implies$ our objectives differentiable w.r.t. T
- If altering x_s (s^{th} covariate) costs γ_s per unit, penalize shift-intervention objective using: $\displaystyle\sum_{s=1}^{d} \gamma_s |\Delta_s|$

  (Use unweighted ℓ_1 penalty find sparse shift interventions, $\gamma_s = 1$)

Algorithms

- Standard GP prior for $f \implies F^{-1}_{G(T)}(\alpha)$ has closed form
- Smooth kernel $\implies$ our objectives differentiable w.r.t. T
- If altering x_s (s^{th} covariate) costs γ_s per unit, penalize shift-intervention objective using: $\sum_{s=1}^{d} \gamma_s |\Delta_s|$

  (Use unweighted ℓ_1 penalty find sparse shift interventions, $\gamma_s = 1$)
- Employ proximal gradient method for optimization of T

Algorithms

- Standard GP prior for $f \implies F^{-1}_{G(T)}(\alpha)$ has closed form
- Smooth kernel $\implies$ our objectives differentiable w.r.t. T
- If altering x_s (s^{th} covariate) costs γ_s per unit, penalize shift-intervention objective using: $\sum_{s=1}^{d} \gamma_s |\Delta_s|$

  (Use unweighted ℓ_1 penalty find sparse shift interventions, $\gamma_s = 1$)
- Employ proximal gradient method for optimization of T
- To avoid poor local maxima, use continuation technique

  (optimize variants of objective with tapering levels of exaggerated smoothness)

Summary of Results

- **Theoretical Guarantee:** As $n \to \infty$: maximizer of our personalized/empirical-population intervention-objectives converges to optimal transformation w.r.t. true f (under reasonable prior)
- **Theoretical Guarantee:** $\forall\ n$: True $f \in$ RKHS of GP prior $\implies$ chosen intervention unlikely to be harmful (probability in terms of α)

Summary of Results

- **Theoretical Guarantee:** As $n \to \infty$: maximizer of our personalized/empirical-population intervention-objectives converges to optimal transformation w.r.t. true f (under reasonable prior)
- **Theoretical Guarantee:** $\forall\ n$: True $f \in$ RKHS of GP prior $\implies$ chosen intervention unlikely to be harmful (probability in terms of α)
- GP-based sparse population intervention outperforms standard frequentist regression methods in gene knockdown application
- Beneficial personalized interventions for writing improvement
  $\alpha = 0.05$ produces far fewer harmful interventions than $\alpha = 0.5$
- Methods work well in misspecified setting (theory + empirical results)
  where sparse-intervention actually affects descendant-covariates in causal DAG

Population Intervention for Gene Perturbation

- X = expression of 10 TF genes[2], Y = expression of small molecule metabolism gene ($n = 161$, try 16 different Y)

[2] Kemmeren P et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* (2014).

Population Intervention for Gene Perturbation

- $X =$ expression of 10 TF genes[2], $Y =$ expression of small molecule metabolism gene ($n = 161$, try 16 different Y)
- Propose single TF knockdown (uniform intervention) which will lead to largest down-regulation of metabolism gene

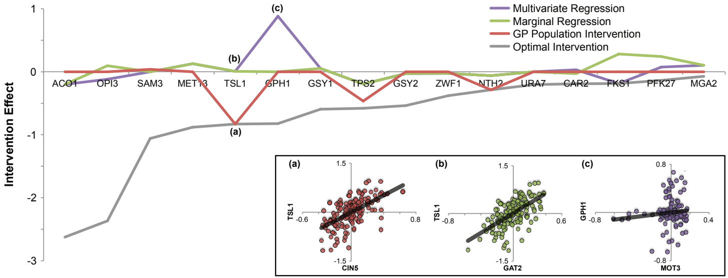
  (verification: single gene deletion applied to each TF in subsequent experiments)

[2] Kemmeren P et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* (2014).

Population Intervention for Gene Perturbation

- X = expression of 10 TF genes[2], Y = expression of small molecule metabolism gene ($n = 161$, try 16 different Y)
- Propose single TF knockdown (uniform intervention) which will lead to largest down-regulation of metabolism gene

  (verification: single gene deletion applied to each TF in subsequent experiments)



[2] Kemmeren P et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* (2014).

Personalized Intervention for Writing Improvement

- $X =$ Various text-features[3] extracted from articles (eg. word-count, polarity, subjectivity), $Y = \#$ of shares on social media $(n = 5000)$

[3] K Fernandes Vinagre P, Cortez P. A proactive intelligent decision support system for predicting the popularity of online news. *EPIA Portuguese Conference on Artificial Intelligence* (2015).

Personalized Intervention for Writing Improvement

- $X =$ Various text-features[3] extracted from articles (eg. word-count, polarity, subjectivity), $Y = \#$ of shares on social media $(n = 5000)$
- Uncertainty-averse method with $\alpha = 0.05$ outperforms alternative which ignores uncertainty ($\alpha = 0.5$), producing half as many harmful interventions without reduction in overall average improvement (evaluated in held-out set of new articles)

[3] K Fernandes Vinagre P, Cortez P. A proactive intelligent decision support system for predicting the popularity of online news. *EPIA Portuguese Conference on Artificial Intelligence* (2015).

Personalized Intervention for Writing Improvement

- $X =$ Various text-features[3] extracted from articles (eg. word-count, polarity, subjectivity), $Y = \#$ of shares on social media $(n = 5000)$
- Uncertainty-averse method with $\alpha = 0.05$ outperforms alternative which ignores uncertainty ($\alpha = 0.5$), producing half as many harmful interventions without reduction in overall average improvement (evaluated in held-out set of new articles)
- Proposes different sparse interventions for articles in Business category vs. Entertainment category: Sparse transformations for business articles uniquely advocate decreasing polarity, whereas interventions to decrease title subjectivity are uniquely prevalent for entertainment articles.

[3] K Fernandes Vinagre P, Cortez P. A proactive intelligent decision support system for predicting the popularity of online news. *EPIA Portuguese Conference on Artificial Intelligence* (2015).

Misspecified Setting

- In practice, sparse interventions may inadvertently affect covariates downstream (in causal DAG) of those chosen for intervention
  (our framework incorrectly assumes T is perfectly enacted)

Misspecified Setting

- In practice, sparse interventions may inadvertently affect covariates downstream (in causal DAG) of those chosen for intervention
  (our framework incorrectly assumes T is perfectly enacted)

- Generate data from underlying non-Gaussian linear structural equation model[4]

[4] Shimizu S et al. A linear non-Gaussian acyclic model for causal discovery. *JMLR* (2006)

Misspecified Setting

- In practice, sparse interventions may inadvertently affect covariates downstream (in causal DAG) of those chosen for intervention
  (our framework incorrectly assumes T is perfectly enacted)
- Generate data from underlying non-Gaussian linear structural equation model[4]
- Find best uniform intervention-policy where T allowed to determine single covariate $s \in \{1, \ldots, d\}$ $(T(x) = [x_1, \ldots, x_{s-1}, z_s, x_{s+1}, \ldots, x_d])$

[4] Shimizu S et al. A linear non-Gaussian acyclic model for causal discovery. *JMLR* (2006)

Misspecified Setting

- In practice, sparse interventions may inadvertently affect covariates downstream (in causal DAG) of those chosen for intervention
  (our framework incorrectly assumes T is perfectly enacted)

- Generate data from underlying non-Gaussian linear structural equation model[4]

- Find best uniform intervention-policy where T allowed to determine single covariate $s \in \{1, \dots, d\}$ $(T(x) = [x_1, \dots, x_{s-1}, z_s, x_{s+1}, \dots, x_d])$

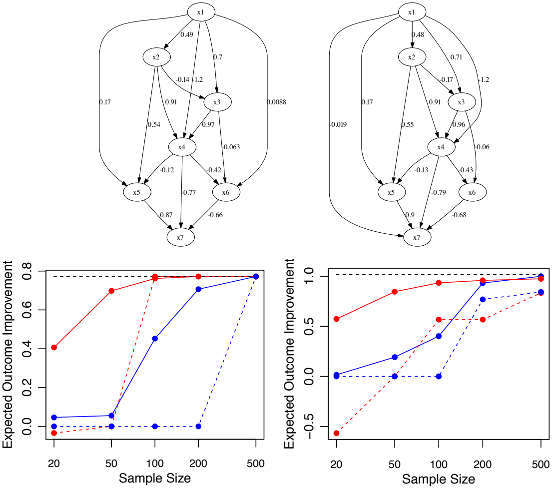
- Intervention actually realized by applying do-operation $do(x_s = z_s)$ in underlying SEM (used to evaluate results)

---

[4] Shimizu S et al. A linear non-Gaussian acyclic model for causal discovery. *JMLR* (2006)

Misspecified Setting

X = [x1, . . . , x6] , Y =x7



Red = uniform intervention selected with GP regression

Blue = best intervention in LinGAM-inferred SEM