Causal inference using invariant prediction: identification and confidence intervals

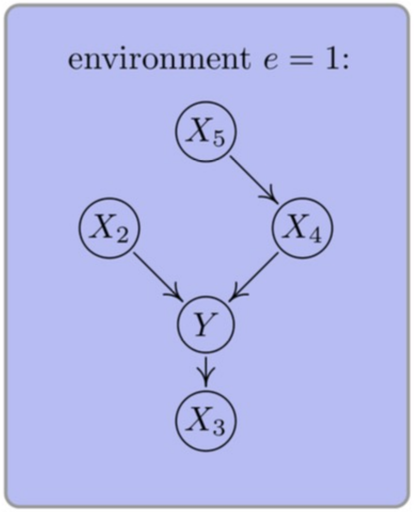
Feb 13th

Zilin Jing

Broad Idea

- Causal Discovery:
  - Discover causal structures given data collected from different environments
- Property: Assume no hidden confounders, target y, all direct parents x.
  - P(y|x) remain identical given any interventions other than y

- Research question: Can we efficiently find a set x2 such that P(y|x2) remain identical. And it is highly possible that x2 is similar to X

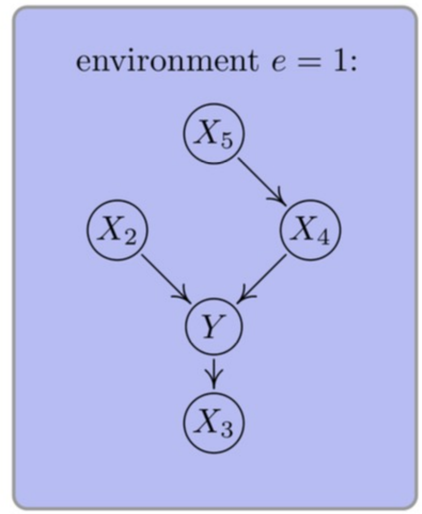


Background – Structural Equation Models

- Linear Gaussian SEMs

Let the first block of data ($e = 1$) always correspond to an "observational" (linear) Gaussian SEM. Here, a distribution over $(X_1^1, \ldots, X_{p+1}^1)$ is said to be generated from a Gaussian SEM if

$$X_j^1 = \sum_{k \neq j} \beta_{j,k}^1 X_k^1 + \varepsilon_j^1, \qquad j = 1, \ldots, p+1, \tag{19}$$



  - Noise variables ε:
  - Variables X
  - Environment: e
- Based on causal graph, we have PA(j), DE(j), AN(j)…

- Different interventions -> Different causal graphs
  - Do-interventions
  - Noise interventions

Invariance Definition

- Assumption1: γ∗ and S* are identical across all environments

Assumption 1 (Invariant prediction) *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

$$\begin{aligned} \text{for all } e \in \mathcal{E}: \quad & X^e \text{ has an arbitrary distribution and} \\ & Y^e = \mu + X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \end{aligned} \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

- Remark:
  - No causality assumption
  - S∗ is not necessarily unique. Consider only one environment
  - P(Y|X) are identical across environments

Relation to causality

- Consider Linear SEMs

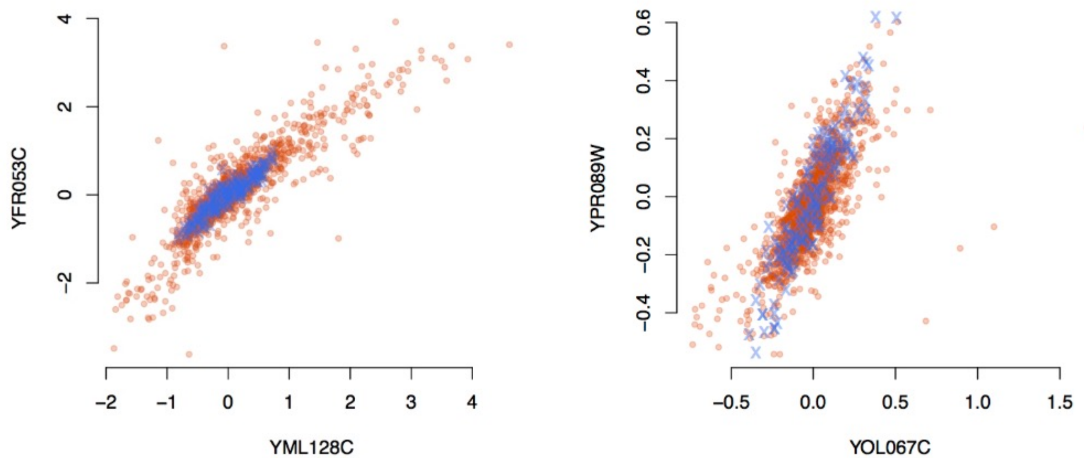
Let the first block of data ($e = 1$) always correspond to an "observational" (linear) Gaussian SEM. Here, a distribution over $(X_1^1, \ldots, X_{p+1}^1)$ is said to be generated from a Gaussian SEM if

$$X_j^1 = \sum_{k \neq j} \beta_{j,k}^1 X_k^1 + \varepsilon_j^1, \qquad j = 1, \ldots, p+1, \tag{19}$$

- All parents of Y form a set S*: $S* = PA(1)$, and $\gamma* = \beta1$
- Proof Sketch:
  - Intervention doesn't influence y or outside noise variable
  - Noise variable independent over Xs (not true with hidden confounders)

Eg: Gene Relation

- If Y|X are identical across different environments?



Plausible Causal Structures

- Motivation: Identify X that satisfy invariance assumption
- Hypothesis test: for each S ⊆ {1,…,p}

$$H_{0,\gamma,S}(\mathcal{E}): \quad \gamma_k = 0 \text{ if } k \notin S \quad \text{and} \quad \begin{cases} \exists F_\varepsilon \text{ such that for all } e \in \mathcal{E} \\ Y^e = X^e\gamma + \varepsilon^e, \text{ where } \varepsilon^e \perp\!\!\!\perp X^e_S \text{ and } \varepsilon^e \sim F_\varepsilon. \end{cases}$$

- Plausible causal predictors

(i) *We call the variables* $S \subseteq \{1,\dots,p\}$ plausible causal predictors *under* $\mathcal{E}$ *if the following null hypothesis holds true:*

$$H_{0,S}(\mathcal{E}): \quad \exists\gamma \in \mathbb{R}^p \text{ such that } H_{0,\gamma,S}(\mathcal{E}) \text{ is true.} \tag{5}$$

(ii) *The* identifiable causal predictors *under interventions* $\mathcal{E}$ *are defined as the following subset of plausible causal predictors*

$$S(\mathcal{E}) := \bigcap_{S:\, H_{0,S}(\mathcal{E}) \text{ is true}} S = \bigcap_{\gamma \in \Gamma(\mathcal{E})} \{k : \gamma_k \neq 0\}. \tag{6}$$

  - Remark: S(E) ⊆ S∗ , S(E1) ⊆ S(E2) if E1 ⊆ E2

Plausible Causal Structures

- Plausible causal coefficients

Definition 2 (Plausible causal coefficients) *We define the set* $\Gamma_S(\mathcal{E})$ *of* plausible causal coefficients for the set $S \subseteq \{1, \dots, p\}$ *and the global set* $\Gamma(\mathcal{E})$ *of* plausible causal coefficients *under* $\mathcal{E}$ *as*

$$\Gamma_S(\mathcal{E}) := \{\gamma \in \mathbb{R}^p : \ H_{0,\gamma,S}(\mathcal{E}) \text{ is true}\}, \quad (7)$$

$$\Gamma(\mathcal{E}) := \bigcup_{S \subseteq \{1,\dots,p\}} \Gamma_S(\mathcal{E}). \quad (8)$$

  - Remark: Γ(E) ∗ ⊆ Γ. Γ(E1) ⊇ Γ(E2) if E1 ⊆ E2.
- Alternative form of H0

$$\beta^{\text{pred},e}(S) := \text{argmin}_{\beta \in \mathbb{R}^p : \beta_k = 0 \text{ if } k \notin S} E(Y^e - X^e \beta)^2$$

$$H_{0,S}(\mathcal{E}) : \quad \begin{cases} \exists \beta \in \mathbb{R}^p \text{ and } \exists F_\varepsilon \text{ such that for all } e \in \mathcal{E} \text{ we have} \\ \beta^{\text{pred},e}(S) \equiv \beta \text{ and } Y^e = X^e \beta + \varepsilon^e, \text{ where } \varepsilon^e \perp\!\!\!\perp X_S^e \text{ and } \varepsilon^e \sim F_\varepsilon. \end{cases} \quad (10)$$

We conclude that

$$\Gamma_S(\mathcal{E}) = \begin{cases} \emptyset & \text{if } H_{0,S}(\mathcal{E}) \text{ is false} \\ \beta^{\text{pred},e}(S) & \text{otherwise.} \end{cases} \quad (11)$$

Construct Good estimators

Generic method for invariant prediction

1) For each set $S \subseteq \{1, \ldots, p\}$, test whether $H_{0,S}(\mathcal{E})$ holds at level α (we will discuss later concrete examples).
2) Set $\hat{S}(\mathcal{E})$ as

$$\hat{S}(\mathcal{E}) := \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ not rejected}} S. \tag{12}$$

3) For the confidence sets, define

$$\hat{\Gamma}(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \hat{\Gamma}_S(\mathcal{E}), \tag{13}$$

where

$$\hat{\Gamma}_S(\mathcal{E}) := \begin{cases} \emptyset & H_{0,S}(\mathcal{E}) \text{ can be rejected at level } \alpha \\ \hat{C}(S) & \text{otherwise.} \end{cases} \tag{14}$$

Here, $\hat{C}(S)$ is a $(1-\alpha)$-confidence set for the regression vector $\beta^{\mathrm{pred}}(S)$ that is obtained by pooling the data.

Good Coverage Guarantee

$$P\big[\hat{S}(\mathcal{E}) \subseteq S^*\big] \geq 1-\alpha. \qquad\qquad P\big[\gamma^* \in \hat{\Gamma}(\mathcal{E})\big] \geq 1-2\alpha.$$

Method1: Regression method

- Observation: For all environments, Regression effects are identical to the causal coefficients

$$\beta^{\text{pred},e}(S^*) \equiv \gamma^* \qquad \text{and} \qquad \sigma^e(S^*) \equiv \text{Var}(F_\varepsilon)^{1/2}.$$

- For each subset, we iterate through all environments
  - Ie be the set of observations in current e, ne = |Ie|. I-e: observations in other environments
  - Train OLS estimator on I-e and generate Yˆe.
  - Compute D := Ye − Yˆe, which follows:

$$\frac{D^t \Sigma_D^{-1} D}{\hat{\sigma}^2 \, n_e} \sim F(n_e, n_{-e} - |S| - 1),$$

  - Reject if p< α/|E|
- Follow generic algorithm to get confidence region for S and γ
- Reject Γ if Γˆ S(E) = ∅, and βpred(S) is:

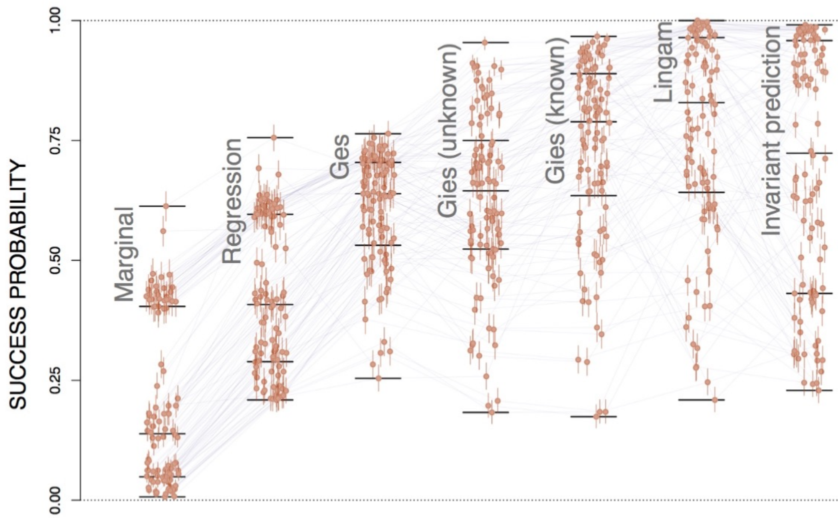
$$(\hat{\beta}^{\text{pred}}(S))_S \pm t_{1-\alpha/(2|S|), n-|S|-1} \cdot \hat{\sigma} \, \text{diag}((\mathbf{X}_S^t \mathbf{X}_S)^{-1}),$$

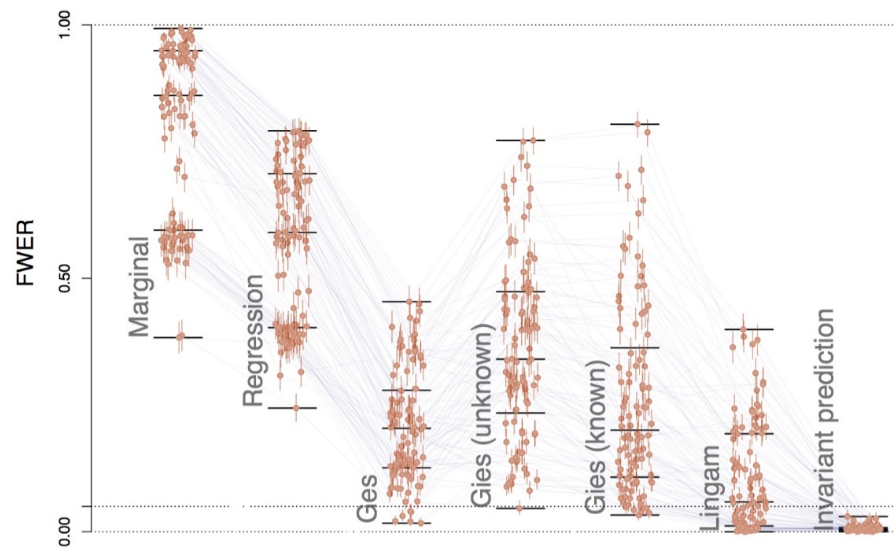
Method2: Faster Approach

- Motivation
  - Avoid computing matrix inversion intensively
  - Extend methods to non-linear approach
- Solution: fit one global model to all data and compare the distribution of the residuals in each experimental setting.
- For each subset, we iterate through all environments
  - Fit a linear regression model on all data to get an estimate β̂pred(S).
  - Compute Residual R = Y − X β̂pred (S) for Re and R-e
  - Subtests:
    - T-test for Mean: H0: E(Re) = E(R-e) -> p value p0_e
    - F test for Variance: H0 Var(Re) = Var(-e) -> p value p1_e
    - Bf correction: Divide each p by |E| and summarize across environments.
    - Test if min{p0,p1} < alpha

Empirical Results - Simulation

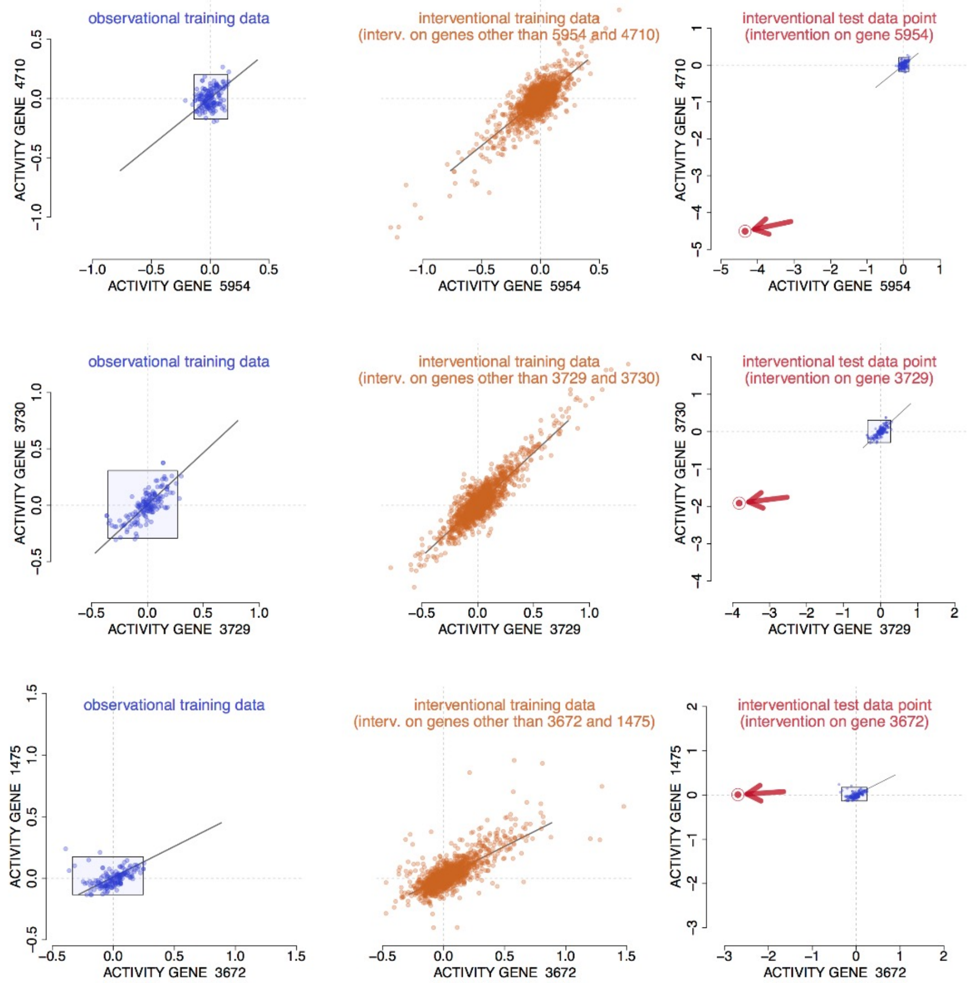
- Data generated by Linear Gaussian SEMs - 100 environment *1000 data
- Test if $\hat{S}(E) = S*$ for each environment
- Baselines: Regression, etc.





Empirical Results – Real Data

- Genes Expression Activities:
  - p = 6170 genes.
  - n_obs = 160, n_int = 1479
  - True positive (x1,x2)
    - X1 is a direct parent of X2, if the activities of x2 intervening after X1 change dramatically (1% upper/lower quantile)



Empirical Results – Real Data

- Method II: eight causal effects that are significant at level 0.01 after a Bonferroni correction

method	Method I	Method II	GIES	IDA	marginal corr.		random guessing
					observ.	pooled	
# of true positives (out of 8)	6	6	2	2	1	2	2 (95% quantile) 3 (99% quantile) 4 (99.9% quantile)

Identifiability results

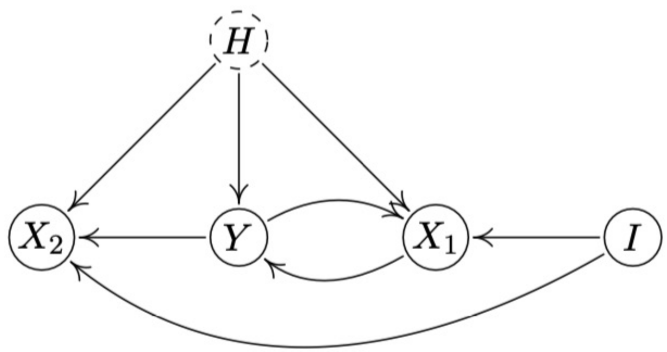
- For a linear Gaussian SCM, Plausible causal predictor always give the true parent

$$S(\mathcal{E}) = \mathbf{PA}(Y) = \mathbf{PA}(1)$$

- Constraint: if interventions are do-interventions, t least one single intervention on each variable other than Y
- We can release the constraint if :
  - Only one intervened environment
  - Let X_k0 be a youngest parent of Y, we intervene on X_k0 is enough

What if hidden variables exists - IV

- Motivation: Hidden variables H exists.



$$X = f(I, H, Y, \eta),$$
$$Y = X\gamma^* + g(H, \varepsilon),$$

  - Regressing Y on X does not yield a consistent estimator for γ∗.
  - Residuals Y – Xs*γ is not always independent of causal predictors Xs
- Def of IV: IV variables only affect Y only through the exposure X and it is independent of confounders H

IV solution

- Solution: Define E as two distinct environments by collecting all samples with I (eg: I=0 vs I=1)
- Construct a weaker hypothesis

$$H_{0,S,hidden}(\mathcal{E}): \quad \exists \gamma \in \mathbb{R}^p \text{ such that } \gamma_k = 0 \text{ if } k \notin S \text{ and}$$
$$\text{the distribution of } Y^e - X^e\gamma \text{ is identical for all } e \in \mathcal{E}.$$

- Estimator

$$\hat{S}(\mathcal{E}) = \bigcap_{S: H_{0,S,hidden}(\mathcal{E}) \text{ not rejected}} S.$$

- Great Coverage

Proposition 2 *Consider model (23) and let* $S^* = \{k : \gamma_k^* \neq 0\}$. *Suppose the test for* $H_{0,S,hidden}(\mathcal{E})$ *is conducted at level* α *and* $\hat{S}$ *is defined as in (26). Then*

$$P[\hat{S}(\mathcal{E}) \subseteq S^*] \geq 1 - \alpha.$$

Q & A