ETH zürich

Invariant Causal Prediction

Causality

Christina Heinze-Deml

Spring 2019

Announcements

- Exam
  - Multiple choice questions
  - No R syntax questions but need to be able to interpret output

- One question hour in July or August

Last week

- Restricted SEMs
  - RESIT
  - LiNGAM

Today

- LiNGAM
- Invariant Causal Prediction

LiNGAM: Linear non-Gaussian acyclic models

- Linear SEM

$$X \leftarrow BX + \epsilon \quad \text{with } B \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^{p}, \epsilon \in \mathbb{R}^{p}$$

- LiNGAM: Linear Non-Gaussian Acyclic Models
  - ϵ is mean-zero non-Gaussian with positive variance
    - Noise components are mutually independent, i.e. no hidden variables (causal sufficiency)
  - No faithfulness assumption needed
  - Estimand: DAG

LiNGAM: Linear non-Gaussian acyclic models

- Linear SEM

$$X \leftarrow BX + \epsilon \quad \text{with } B \in \mathbb{R}^{p \times p}, X \in \mathbb{R}^p, \epsilon \in \mathbb{R}^p$$

- Due to acyclicity, the diagonal elements of B are zero
  - No self-loops, i.e. no edge from a node to itself

- Permuting the order of the variables using a causal ordering makes B strictly lower triangular
  - I.e. due to acyclicity, always possible to perform simultaneous, equal row and column permutations on B to make it strictly lower triangular

Independent component analysis

- Independent component analysis (ICA)
  - ICA model

$$X = AS$$

    - $X \in \mathbb{R}^p$: observed variables
    - $S \in \mathbb{R}^p$: mutually independent, continuous latent non-Gaussian variables – “sources”
    - $A \in \mathbb{R}^{p \times p}$: unobserved full-rank mixing matrix

  - If S is non-Gaussian, then A is identifiable up to permutation, scaling and sign of the columns

LiNGAM: Linear non-Gaussian acyclic models

- Can write:

$$X = BX + \epsilon$$

$$(I - B)X = \epsilon$$

$$X = (I - B)^{-1}\epsilon$$

- LiNGAM is an instance of the ICA model $X = AS$ with $A = (I - B)^{-1}$ and $S = \epsilon$
- Recall: A is identifiable up to permutation, scaling and sign of the columns
  - Can exploit further properties of B: “zeros on the diagonal” and “strictly lower triangular”

Example

LiNGAM: Linear non-Gaussian acyclic models

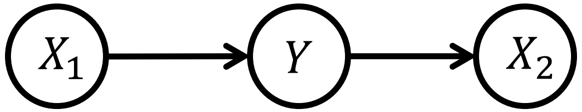
- ICA-LiNGAM algorithm:
  1. Given n i.i.d. observations of $X_{\mathbf{V}}$, use ICA to estimate $\mathrm{W} = A^{-1} = (I - B)$ up to permutation, scaling and sign of the columns
  2. Find unique permutation of the rows of W that yields $\widetilde{\mathrm{W}}$ without any zeros on the diagonal
     - Permutation is found by minimizing $\sum_i 1/|\widetilde{W}_{ii}|$ (classical linear assignment problem)
  3. Divide each row of $\widetilde{\mathrm{W}}$ by its diagonal element to yield $\widetilde{\mathrm{W}}'$ with only ones on the diagonal
  4. Compute $\hat{B} = I - \widetilde{\mathrm{W}}'$
  5. Find causal order by making $\tilde{B} = \tilde{P}\hat{B}\tilde{P}^T$ as close as possible to strictly lower triangular
     - Prune edge weights, e.g. using sparse regression or significance testing

Shimizu et al. (2006)

Invariant causal prediction

- Knowing causal structure can help to improve predictions when the underlying distribution changes
  - Example:

$$X_1 \leftarrow N_{X_1}$$
$$Y \leftarrow X_1 + N_Y$$
$$X_2 \leftarrow Y + N_{X_2}$$



with $N_{X_1}, N_Y \sim \mathcal{N}(0,1)$ and $N_{X_2} \sim \mathcal{N}(0, 0.1)$, jointly independent

  - Interested in predicting Y from X_1 and X_2
  - Model 1: linear model $Y \sim X_1$
  - Model 2: linear model $Y \sim X_2$
  - MSE of model 2 smaller than MSE of model 1
  - If interventions occur on X_1 and X_2 after fitting the model, model 1 still works while model 2 fails

Invariant causal prediction

- Knowing causal structure can help to improve predictions when the underlying distribution changes
- Now: Observing a system in different "environments" can be used to learn causal relations

- Setting: Assume we observe data from different environments $e \in \mathcal{E}$:

$$X_{\boldsymbol{V}}^e = (X_1^e, \ldots, X_d^e) \sim P^e$$

i.e., each variable X_j is measured in different environments $e \in \mathcal{E}$

Invariant causal prediction

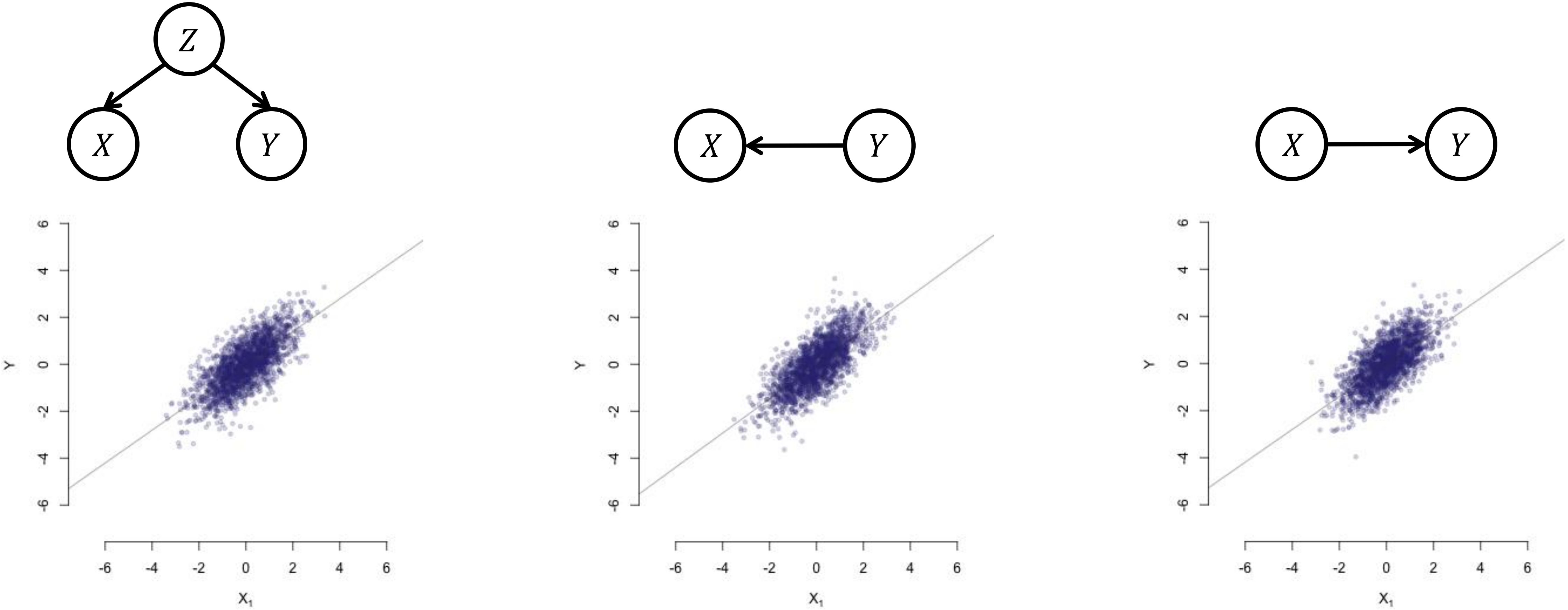
- Consider target variable Y with a set of p predictor variables $(X_1, \dots, X_p)$

- Goal: instead of learning the whole causal graph, find causal parents of Y
  - Denote the set of causal parents of Y by S^*, i.e. $S^* \coloneqq \mathrm{pa}(Y)$

- Both $(X_1, \dots, X_p)$ and Y are observed in different environments $e \in \mathcal{E}$
  - These different environments could be intervention settings with unknown targets
  - Can also exploit time series data

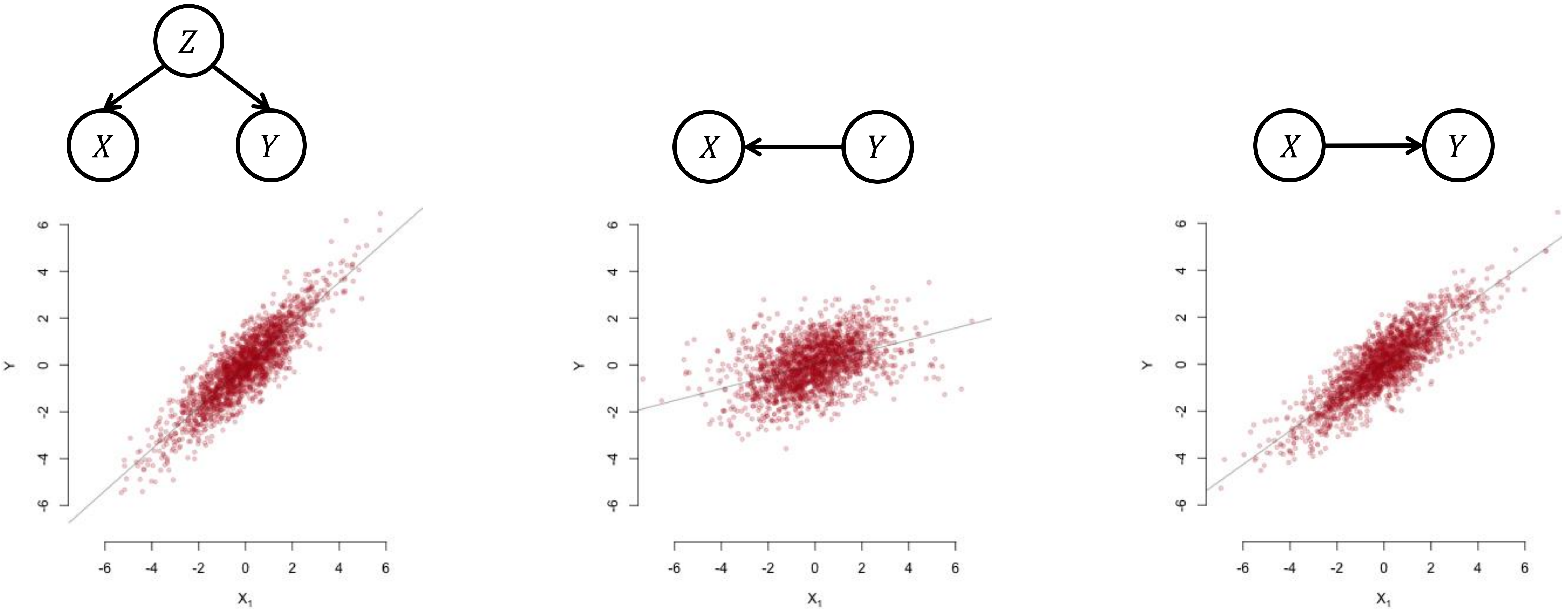
Invariance

- Many SEMs generate the same observational distribution



Invariance

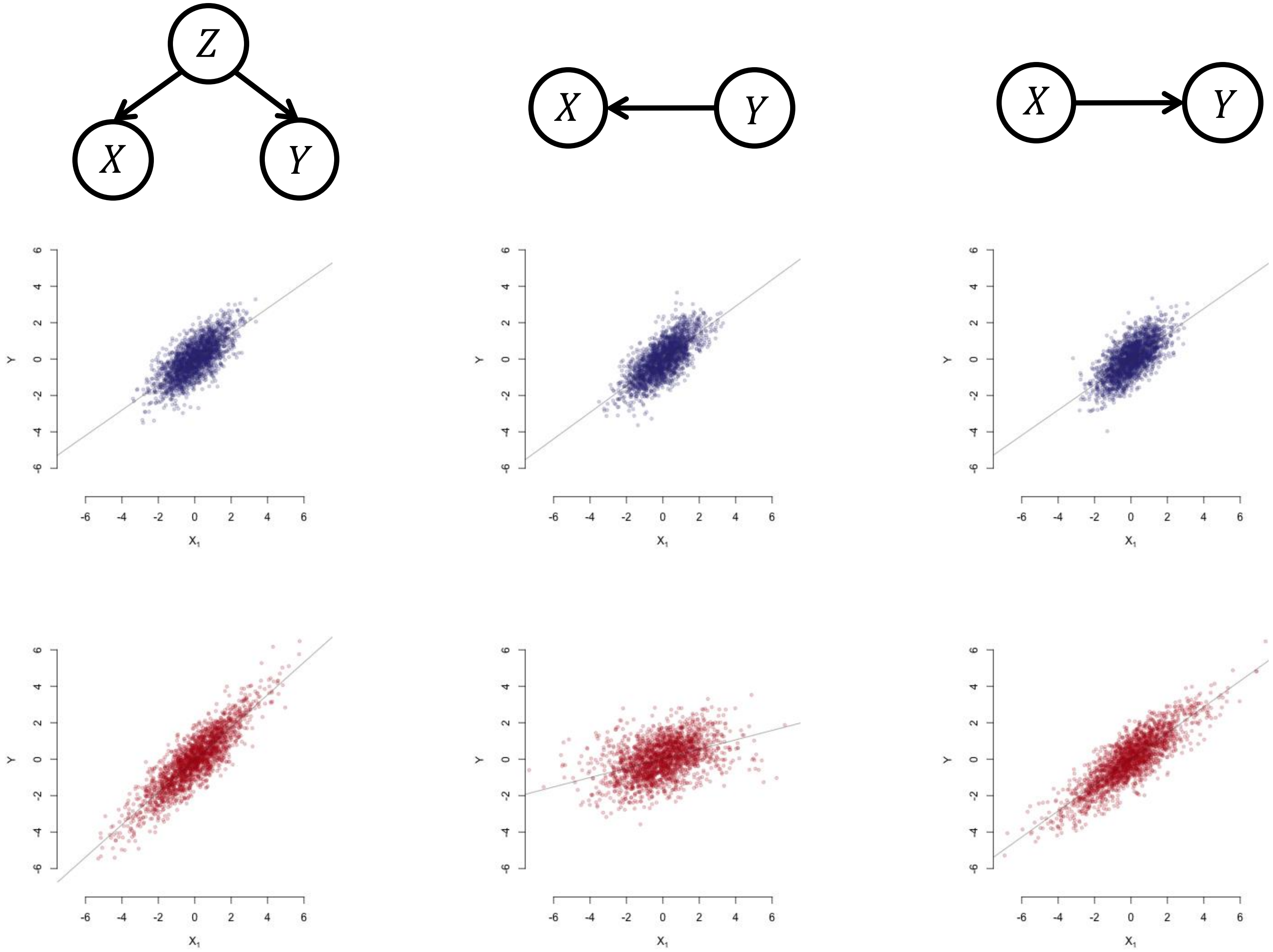
- Now, we observe the data under shift interventions on X



Invariance

Observational data
Environment 1

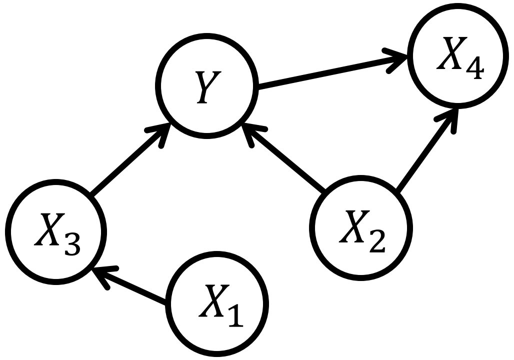
Interventional data
Environment 2



Invariance

- As long as we avoid interventions on the target Y itself, for all environments $e \in \mathcal{E}$:

$$\begin{cases} X^e \text{ has an arbitrary distribution} \\ Y^e = f_Y(X^e_{S^*}, N_Y) \end{cases}$$



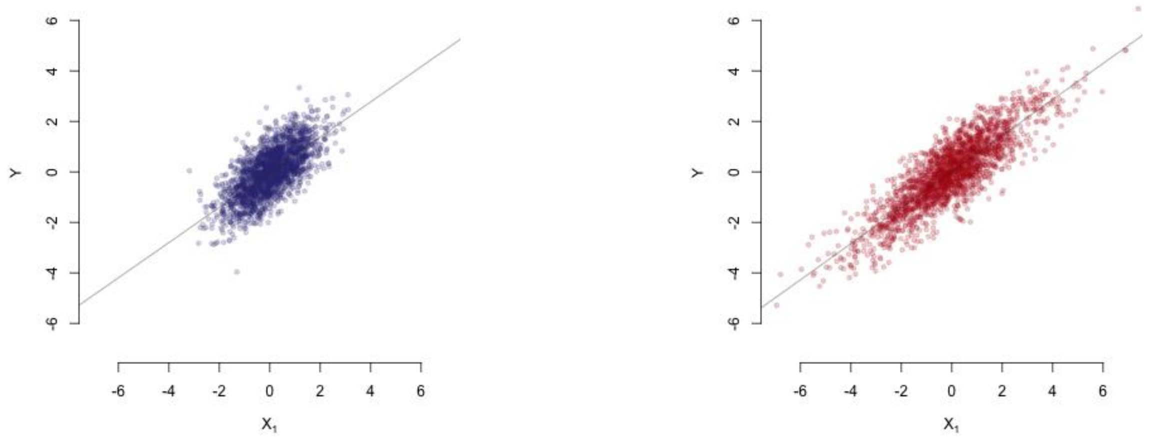
- For all environments $e, f \in \mathcal{E}$

$$Y^e | X^e_{S^*} = x \quad = \quad Y^f | X^f_{S^*} = x$$

Invariant causal prediction

- Idea: find invariant conditional distributions to estimate $S^* \coloneqq \mathrm{pa}(Y)$

- No search over DAG space necessary
- No faithfulness assumption necessary
- Relies on data from multiple environments



Peters, Bühlmann, Meinshausen (2016)

Invariant causal prediction for linear models

- Let S^* be the indices of $\mathrm{pa}(Y)$.

  $H_{0,S^*}(\mathcal{E})$: There exists $\gamma^* \in \mathbb{R}^p$ with support S^* that satisfies for all $e \in \mathcal{E}$:

  X^e has an arbitrary distribution and

  $Y^e = X^e\gamma^* + \epsilon^e, \epsilon^e \sim F_\epsilon$ and $\epsilon^e \perp\!\!\!\perp X^e_{S^*}$

- Idea: To find S^*, test null hypothesis $H_{0,S}(\mathcal{E})$ for all subsets of the predictor
- Estimate: $\hat{S} := \cap_{S:H_{0,S} \text{ not rej.}} S$

Guarantee

- Idea: To find S^*, test null hypothesis $H_{0,S}(\mathcal{E})$ for all subsets of the predictor
- Estimate: $$\hat{S} := \cap_{S:H_{0,S} \text{ not rej.}} S$$
- Theorem: “no mistakes”:

$$P\left(\hat{S} \subseteq S^*\right) \geq P(H_{0,S^*} \text{ not rejected}) \geq 1 - \alpha$$

Invariant Causal Prediction

- Idea: To find S^*, test null hypothesis $H_{0,S}(\mathcal{E})$ for all subsets of the predictor
- How to formulate $H_{0,S}(\mathcal{E})$?
  - Different options
  - Approximate test on residuals:

    For each $S \subseteq \{1, \dots, p\}$:

    - Fit linear regression using set S of variables and data from all environments. Let $R = Y - \hat{f}(X_S)$.
    - Test the null hypotheses that the means and the variances of R are identical across all environments. Combine the two *p*-values by taking twice the smaller of the two values.
    - If the combined *p*-value is smaller than α, reject the set S.

Recap

- Concepts to know:
  - Invariant Causal Prediction

References and acknowledgments

- Peters, Janzing and Schölkopf (2017). Elements of Causal Inference. Chapters 7.1, 7.2
- Invariant Causal Prediction
  - Peters, Bühlmann, Meinshausen (2016). Causal inference using invariant prediction: identification and confidence intervals. JRSS B.