Neural Causal AI

Adversarial Invariance Learning from Multiple Environments

Jianqing Fan

Princeton University

with **Yihong Gu**, **Cong Fang**, and **Peter Buehlmann**

Outlines

1. Introduction
2. Endogeneity in High Dimension
3. Multi-Environment Linear Regression
4. Neural Causal Learning
5. Causality under SCM
6. Implementation and Numerical Studies

Neural Causai AI

Outlines

1. Introduction
2. Endogeneity in High Dimension
3. Multi-Environment Linear Regression
4. Neural Causal Learning
5. Causality under SCM
6. Implementation and Numerical Studies

Neural Causai AI

Yihong Gu

Cong Fang

Peter Buelhmann

Outlines

1. Introduction
2. Endogeneity in High Dimension
3. Multi-Environment Linear Regression
4. Neural Causal Learning
5. Causality under SCM
6. Implementation and Numerical Studies

Neural Causai AI

Yihong Gu

Cong Fang

Peter Buelhmann

Introduction

What Is Causality?

Wikipedia: one event, process, or state contributes to **production of another**:

- ★ relations hold in the past must hold in the future
- ★ relations hold in one environment must hold in another.

Invariance

Phil. of Sci.: Phenomenon that no evidence against is regarded a truth.

Causality $\approx$ Invariance under MEs

What Is Causality?

Wikipedia: one event, process, or state contributes to **production of another**:

- ★ relations hold in the past must hold in the future
- ★ relations hold in one environment must hold in another.

Invariance

Phil. of Sci.: Phenomenon that no evidence against is regarded a truth.

Causality $\approx$ Invariance under MEs

What Is Causality?

Wikipedia: one event, process, or state contributes to **production of another**:

- ★ relations hold in the past must hold in the future
- ★ relations hold in one environment must hold in another.

Invariance

Phil. of Sci.: Phenomenon that no evidence against is regarded a truth.

Causality $\approx$ Invariance under MEs

What Is Causality?

Wikipedia: one event, process, or state contributes to **production of another**:

- ★ relations hold in the past must hold in the future
- ★ relations hold in one environment must hold in another.

Invariance

Phil. of Sci.: Phenomenon that no evidence against is regarded a truth.

Causality $\approx$ Invariance under MEs

Statistical Learning

★Prediction ★Attribution ★Inferences ★Causality

Typical Processes:

- ★ Collect response variable Y and its associated variables $X \in \mathbb{R}^p$.
- ★ Use statistical machine algorithms to select important variables.

What can be wrong?

Statistical Learning

★Prediction ★Attribution ★Inferences ★Causality

Typical Processes:

- ★ Collect response variable Y and its associated variables $X \in \mathbb{R}^p$.
- ★ Use statistical machine algorithms to select important variables.

What can be wrong?

An illustrative example

★ **Classification uses two features:**

body shape $X_1 \longrightarrow$

background color $X_2 \longrightarrow$

$\boxed{\phi(\cdot)}$

0 : cow 

1 : camel 

★ **Standard SML** : Data $\mathcal{D}$: •70% cows on grass (X_2 green),

•80% camels on sand (X_2 yellow)

Get $\mathcal{D}_{train} + \mathcal{D}_{test}$ $\Longrightarrow$ $\widehat{\phi}(\cdot)$ works well on $\mathcal{D}_{test}$,

but relies on X_2 (spurious)

★ **Prediction**: Not robust in other environments (marketing).

★ **Attribution**: Wrong mechanism or treatment targets!

An illustrative example

★ **Classification uses two features:**

body shape $X_1 \longrightarrow$

background color $X_2 \longrightarrow$

$\phi(\cdot)$ $\longrightarrow$

0 : cow

1 : camel

★ **Standard SML** : Data $\mathcal{D}$: •70% cows on grass (X_2 green),

•80% camels on sand (X_2 yellow)

Get $\mathcal{D}_{train} + \mathcal{D}_{test}$ $\Longrightarrow$ $\widehat{\phi}(\cdot)$ works well on $\mathcal{D}_{test}$,

but relies on X_2 (spurious)

★ **Prediction**: Not robust in other environments (marketing).

★ **Attribution**: Wrong mechanism or treatment targets!

An illustrative example

★ **Classification uses two features:**

body shape $X_1 \longrightarrow$

background color $X_2 \longrightarrow$

$\phi(\cdot)$

$\longrightarrow$ 0 : cow

$\longrightarrow$ 1 : camel

★ **Standard SML** : Data $\mathcal{D}$: •70% cows on grass (X_2 green),

•80% camels on sand (X_2 yellow)

Get $\mathcal{D}_{train} + \mathcal{D}_{test}$ $\Longrightarrow$ $\widehat{\phi}(\cdot)$ works well on $\mathcal{D}_{test}$, but relies on X_2 (spurious)

What is wrong?

★ **Prediction**: Not robust in other environments (marketing).

★ **Attribution**: Wrong mechanism or treatment targets!

An illustrative example

★ **Classification uses two features:**

body shape $X_1 \longrightarrow$

background color $X_2 \longrightarrow$

$\boxed{\phi(\cdot)} \longrightarrow$

0 : cow

1 : camel

★ **Standard SML** : Data $\mathcal{D}$: •70% cows on grass (X_2 green),

•80% camels on sand (X_2 yellow)

Get $\mathcal{D}_{train} + \mathcal{D}_{test} \quad \Longrightarrow \quad \widehat{\phi}(\cdot)$ works well on $\mathcal{D}_{test}$,

but relies on X_2 (spurious)

★ **Prediction**: Not robust in other environments (marketing).

★ **Attribution**: Wrong mechanism or treatment targets!

An illustrative example

★ **Classification uses two features:**

body shape $X_1 \longrightarrow$

background color $X_2 \longrightarrow$

$\boxed{\phi(\cdot)}$ → 0 : cow / 1 : camel

★ **Standard SML** : Data $\mathcal{D}$: •70% cows on grass (X_2 green),

•80% camels on sand (X_2 yellow)

Get $\mathcal{D}_{train} + \mathcal{D}_{test}$ $\Longrightarrow$ $\widehat{\phi}(\cdot)$ works well on $\mathcal{D}_{test}$, but relies on X_2 (spurious)

★ **Prediction**: Not robust in other environments (marketing).

★ **Attribution**: Wrong mechanism or treatment targets!

Can Machine Learn Causality?

Eliminate endogeneity?

Train a Causal AI + What Data?

Questions

Can Machine Learn Causality?

Eliminate endogeneity?

Train a Causal AI + What Data?

Questions

Can Machine Learn Causality?

Eliminate endogeneity?

Train a Causal AI + What Data?

Use data heterogeneity

$\mathcal{D}$: 70% cows on grass
80% camels on sand

$\tilde{\mathcal{D}}$: 50% cows on grass
60% camels on sand

assoc. of X_2 and Y varies in $\mathcal{D}$ and $\tilde{\mathcal{D}} \Longrightarrow X_2$ is spurious variable

X_2 endogenous spurious $\Longrightarrow$ inconsistency

Today's Talk: Variable Selection (Causality Learning) from **Invariance**

X_1 ✓ , X_2 × (X_1, X_2) × X_3 =temperature ✓

Use data heterogeneity

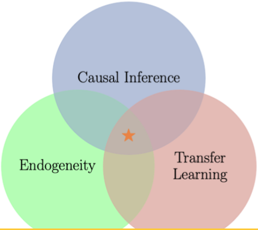
$\mathcal{D}$: 70% cows on grass
80% camels on sand

$\tilde{\mathcal{D}}$: 50% cows on grass
60% camels on sand

assoc. of X_2 and Y varies in $\mathcal{D}$ and $\tilde{\mathcal{D}} \Longrightarrow X_2$ is spurious variable

X_2 **endogenous spurious** $\Longrightarrow$ **inconsistency**

Today's Talk: Variable Selection (Causality Learning) from **Invariance**



Use data heterogeneity

$\mathcal{D}$: 70% cows on grass
80% camels on sand

$\tilde{\mathcal{D}}$: 50% cows on grass
60% camels on sand

assoc. of X_2 and Y varies in $\mathcal{D}$ and $\tilde{\mathcal{D}} \Longrightarrow X_2$ is spurious variable

X_2 **endogenous spurious** $\Longrightarrow$ **inconsistency**

Today's Talk: Variable Selection (Causality Learning) from **Invariance**

$X_1 \checkmark$, $X_2 \times$ $(X_1, X_2) \times$ $X_3 =$temperature $\checkmark$

Use data heterogeneity

$\mathcal{D}$: 70% cows on grass
80% camels on sand

$\tilde{\mathcal{D}}$: 50% cows on grass
60% camels on sand

assoc. of X_2 and Y varies in $\mathcal{D}$ and $\tilde{\mathcal{D}} \Longrightarrow X_2$ is spurious variable

X_2 endogenous spurious $\Longrightarrow$ inconsistency

Today's Talk: Variable Selection (Causality Learning) from **Invariance**

X_1 ✓, X_2 ✗ (X_1, X_2) ✗ X_3 =temperature ✓

Use data heterogeneity

$\mathcal{D}$: 70% cows on grass
80% camels on sand

$\tilde{\mathcal{D}}$: 50% cows on grass
60% camels on sand

assoc. of X_2 and Y varies in $\mathcal{D}$ and $\tilde{\mathcal{D}} \Longrightarrow X_2$ is spurious variable

X_2 **endogenous spurious** $\Longrightarrow$ **inconsistency**

Today's Talk: Variable Selection (Causality Learning) from **Invariance**

X_1 ✓, X_2 × (X_1, X_2) × X_3 =temperature ✓

Use data heterogeneity

$\mathcal{D}$: 70% cows on grass
80% camels on sand

$\tilde{\mathcal{D}}$: 50% cows on grass
60% camels on sand

assoc. of X_2 and Y varies in $\mathcal{D}$ and $\tilde{\mathcal{D}} \Longrightarrow X_2$ is spurious variable

X_2 **endogenous spurious** $\Longrightarrow$ **inconsistency**

Today's Talk: Variable Selection (Causality Learning) from **Invariance**

X_1 ✓ , X_2 × (X_1, X_2) × X_3 =temperature ✓

exogeneous spurious

Spurious variables

$$\begin{cases} \textbf{Endogeneous:} \\ \textbf{Exogeneous:} \end{cases}$$

Spurious variables

$$\begin{cases} \textbf{Endogeneous:} & \text{background colors} \\ \textbf{Exogeneous:} & \text{time photo taken} \end{cases}$$

Spurious variables

$$\begin{cases} \textbf{Endogeneous:} & \text{background colors} & \textbf{harmful–bias} \\ \textbf{Exogeneous:} & \text{time photo taken} & \textbf{unbiased–var.} \end{cases}$$

ℓ_1, SCAD, SIS

Spurious variables

$$\begin{cases} \textbf{Endogeneous:} & \text{background colors} & \textbf{harmful–bias} \\ \textbf{Exogeneous:} & \text{time photo taken} & \textbf{unbiased–var.} \end{cases}$$

ℓ_1, SCAD, SIS

Elliminate endogeneity by FAIR-NN

Spurious variables

$$\begin{cases} \textbf{Endogeneous:} & \text{background colors} & \textbf{harmful–bias} \\ \textbf{Exogeneous:} & \text{time photo taken} & \textbf{unbiased–var.} \end{cases}$$

ℓ_1, SCAD, SIS

Elliminate endogeneity by FAIR-NN

What only one environment?

Endogeneity in High Dimension

★ Fan, J. and Liao, Y. (2014). Endogeneity in ultrahigh dimension. *Ann. Statist.*, **42**, 872-917.

★ Fan, J., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *Natl. Sci. Rev.*, **1**, 293-314.

Assumptions in Variable Selection

Stylized Model: $Y = \mathbf{X}^T \beta_0 + \varepsilon$, $\mathbb{E}\varepsilon\mathbf{X} = 0$ **or stronger**, β_0 **sparse**.

★Tons of equations! ★ can not be validated!

Assumptions in Variable Selection

Stylized Model: $Y = \mathbf{X}^T\beta_0 + \varepsilon$, $\mathbb{E}\varepsilon\mathbf{X} = 0$ **or stronger**, β_0 **sparse**.

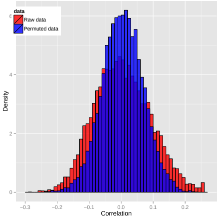
★Tons of equations! ★ $\boxed{\text{can not be validated!}}$

Prostate cancer study

Data: 148 microarrays from GEO data

Response: Expressions of gene DDR

Covariates: remaining 12,718 genes



Assumptions in Variable Selection

Stylized Model: $Y = \mathbf{X}^T\beta_0 + \varepsilon$, $\mathbb{E}\varepsilon\mathbf{X} = 0$ **or stronger**, β_0 **sparse**.

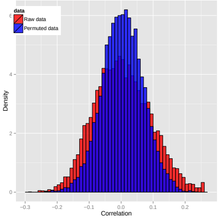
★Tons of equations! ★ $\boxed{\text{can not be validated!}}$

Prostate cancer study

Data: 148 microarrays from GEO data

Response: Expressions of gene DDR

Covariates: remaining 12,718 genes



Example: $Y = 2X_1 + X_2 + \varepsilon$, $\mathbb{E}(\varepsilon|X_1) = 0, \mathbb{E}(\varepsilon|X_2) = 0$

Assumptions in Variable Selection

Stylized Model: $Y = \mathbf{X}^T \beta_0 + \varepsilon$, $\mathbb{E}\varepsilon\mathbf{X} = 0$ **or stronger**, β_0 **sparse**.

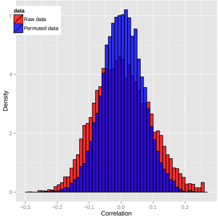
★Tons of equations! ★ can not be validated!

Prostate cancer study

Data: 148 microarrays from GEO data

Response: Expressions of gene DDR

Covariates: remaining 12,718 genes



Example: $Y = 2X_1 + X_2 + \varepsilon$, $\mathbb{E}(\varepsilon|X_1) = 0, \mathbb{E}(\varepsilon|X_2) = 0$

Netting: Collecting many variables $\{X_j\}_{j=1}^p$.

Assumptions in Variable Selection

Stylized Model: $Y = \mathbf{X}^T\beta_0 + \varepsilon$, $\mathbb{E}\varepsilon\mathbf{X} = 0$ **or stronger**, β_0 **sparse**.

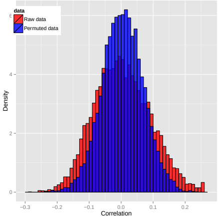
★Tons of equations! ★ $\boxed{\text{can not be validated!}}$

Prostate cancer study

Data: 148 microarrays from GEO data

Response: Expressions of gene DDR

Covariates: remaining 12,718 genes



Example: $Y = 2X_1 + X_2 + \varepsilon$, $\mathbb{E}(\varepsilon|X_1) = 0, \mathbb{E}(\varepsilon|X_2) = 0$

Netting: Collecting many variables $\{X_j\}_{j=1}^p$.

■Many X_j's related to Y,

Assumptions in Variable Selection

Stylized Model: $Y = \mathbf{X}^T\beta_0 + \varepsilon$, $\mathbb{E}\varepsilon\mathbf{X} = 0$ **or stronger**, β_0 **sparse**.

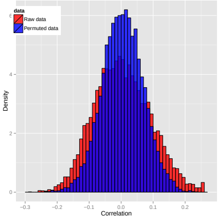
★Tons of equations! ★ can not be validated!

Prostate cancer study

Data: 148 microarrays from GEO data

Response: Expressions of gene DDR

Covariates: remaining 12,718 genes



Example: $Y = 2X_1 + X_2 + \varepsilon$, $\mathbb{E}(\varepsilon|X_1) = 0, \mathbb{E}(\varepsilon|X_2) = 0$

Netting: Collecting many variables $\{X_j\}_{j=1}^p$.

■Many X_j's related to Y, hence to $\varepsilon = Y - 2X_1 - X_2$ for **large** p:

$$\mathrm{corr}(X_j, \varepsilon) \neq 0, \text{for some } j.$$ **Endogeneity**

Solutions

Model: $Y = \mathbf{X}_{S_0}^T \beta_{S_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{S_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{S_0} = 0$,

Solutions

Model: $Y = \mathbf{X}_{S_0}^T \beta_{S_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{S_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{S_0} = 0$,

Solutions

Model: $Y = \mathbf{X}_{\mathcal{S}_0}^T \beta_{\mathcal{S}_0} + \varepsilon$ with $\mathbb{E}(\varepsilon | \mathbf{X}_{\mathcal{S}_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon \mathbf{X}_{\mathcal{S}_0} = 0$, $\mathbb{E}\varepsilon \mathbf{X}_{\mathcal{S}_0}^2 = 0$,

★Variables $\mathbf{X}_{\mathcal{S}_0}$ are special or causal, as more equations than unknowns.

Solutions

Model: $Y = \mathbf{X}_{\mathcal{S}_0}^T \beta_{\mathcal{S}_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{\mathcal{S}_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{\mathcal{S}_0} = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{\mathcal{S}_0}^2 = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{\mathcal{S}_0}^3 = 0$

★Variables $\mathbf{X}_{\mathcal{S}_0}$ are special or causal, as more equations than unknowns. **Invariant**

Solutions

Model: $Y = \mathbf{X}_{S_0}^T \beta_{S_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{S_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{S_0} = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{S_0}^2 = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{S_0}^3 = 0$

★Variables $\mathbf{X}_{S_0}$ are special or causal, as more equations than unknowns. **Invariant**

Generalization: $\boxed{\mathbb{E}(Y - \mathbf{X}_{S_0}^T \beta_{S_0}) f(\mathbf{X}_{S_0}) = 0 \text{ for } f \in \mathcal{F}}$ **GM constraints**

Solutions

Model: $Y = \mathbf{X}_{S_0}^T \beta_{S_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{S_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{S_0} = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{S_0}^2 = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{S_0}^3 = 0$

★Variables $\mathbf{X}_{S_0}$ are special or causal, as more equations than unknowns. **Invariant**

Generalization: $\boxed{\mathbb{E}(Y - \mathbf{X}_{S_0}^T \beta_{S_0}) f(\mathbf{X}_{S_0}) = 0 \text{ for } f \in \mathcal{F}}$ **GM constraints**

Constrained LS

Solutions

Model: $Y = \mathbf{X}_{S_0}^T \beta_{S_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{S_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{S_0} = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{S_0}^2 = 0$,

★Variables $\mathbf{X}_{S_0}$ are special or causal, as more equations than unknowns. **Invariant**

Generalization: $\boxed{\mathbb{E}(Y - \mathbf{X}_{S_0}^T\beta_{S_0})f(\mathbf{X}_{S_0}) = 0 \text{ for } f \in \mathcal{F}}$ **GM constraints**

Soft Constrained LS

$$\min_{\beta} \sum_{i=1}^n \varepsilon_i^2 + \boxed{\lambda(\|\textstyle\sum_{i=1}^n \varepsilon_i \mathbf{X}_{i,S_0}\|^2 + \|\sum_{i=1}^n \varepsilon_i \mathbf{X}_{i,S_0}^2\|^2)}, \qquad \varepsilon_i = Y_i - \mathbf{X}_{i,S_0}^T\beta_{S_0}.$$

Solutions

Model: $Y = \mathbf{X}_{\mathcal{S}_0}^T \beta_{\mathcal{S}_0} + \varepsilon$ with $\mathbb{E}(\varepsilon|\mathbf{X}_{\mathcal{S}_0}) = 0$ or **weaker**. **more realstic**

Example: $\mathbb{E}\varepsilon\mathbf{X}_{\mathcal{S}_0} = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{\mathcal{S}_0}^2 = 0$, $\mathbb{E}\varepsilon\mathbf{X}_{\mathcal{S}_0}^3 = 0$

★Variables $\mathbf{X}_{\mathcal{S}_0}$ are special or causal, as more equations than unknowns. **Invariant**

Generalization: $\boxed{\mathbb{E}(Y - \mathbf{X}_{\mathcal{S}_0}^T \beta_{\mathcal{S}_0}) f(\mathbf{X}_{\mathcal{S}_0}) = 0 \text{ for } f \in \mathcal{F}}$ **GM constraints**

Soft Constrained LS

$$\min_{\beta} \sum_{i=1}^{n} \varepsilon_i^2 + \boxed{\lambda\left(\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(\mathbf{X}_{i,\mathcal{S}_0})\right)}, \qquad \varepsilon_i = Y_i - \mathbf{X}_{i,\mathcal{S}_0}^T \beta_{\mathcal{S}_0}$$

Multi-Environment Linear Reg

★ Fan, J., Fang, C., Gu, Y., and Zhang, T. (2024+). Environment Invariant Linear Least Squares. *Ann. Statist.*

Model

★ **Multi-environment regression**: For each e, $\left(X_i^{(e)}, Y_i^{(e)}\right)_{i=1}^{n} \sim_{i.i.d.} \mu^{(e)} \in \mathcal{U}_{\beta^*}$:

$$\boxed{Y^{(e)} = (\beta^*_{S^*})^\top X^{(e)}_{S^*} + \varepsilon^{(e)}} \quad \text{with} \quad \mathbb{E}[\varepsilon^{(e)} X^{(e)}_{S^*}] = 0.$$

- ♦ $\boxed{S^*, \beta^*}$ are **invariant**.
- ♦ More realistic and weaker than $\mathbb{E}[\varepsilon^{(e)} X^{(e)}] = 0$ for regression.

★ **Heterogeneous**: Each environment does not provide a consistent estimator.

Model

★ **Multi-environment regression**: For each e, $(X_i^{(e)}, Y_i^{(e)})_{i=1}^{n} \sim_{i.i.d.} \mu^{(e)} \in \mathcal{U}_{\beta^*}$:

$$\boxed{Y^{(e)} = (\beta^*_{S^*})^\top X^{(e)}_{S^*} + \varepsilon^{(e)}} \quad \text{with} \quad \mathbb{E}[\varepsilon^{(e)} X^{(e)}_{S^*}] = 0.$$

- $\boxed{S^*, \beta^*}$ are **invariant**. ← $\boxed{\text{learning object}}$
- More realistic and weaker than $\mathbb{E}[\varepsilon^{(e)} X^{(e)}] = 0$ for regression.

★ **Heterogeneous**: Each environment does not provide a consistent estimator.

Model

★ **Multi-environment regression**: For each e, $\left(X_i^{(e)}, Y_i^{(e)}\right)_{i=1}^{n} \sim_{i.i.d.} \mu^{(e)} \in \mathcal{U}_{\beta^*}$:

$$\boxed{Y^{(e)} = (\beta^*_{\mathbf{S}^*})^\top X^{(e)}_{\mathbf{S}^*} + \varepsilon^{(e)}} \quad \text{with} \quad \mathbb{E}[\varepsilon^{(e)} X^{(e)}_{\mathbf{S}^*}] = 0.$$

- ♦ $\boxed{S^*, \beta^*}$ are **invariant**. ← learning object
- ♦ More realistic and weaker than $\mathbb{E}[\varepsilon^{(e)} X^{(e)}] = 0$ for regression.

★ **Heterogeneous**: Each environment does not provide a consistent estimator.

Focused linear invariance regularizer

Population-level penalty:

★delete endogenous variables

$$J(\beta) = \sum_{j \in \mathcal{S}(\beta)} \sum_{e=1}^{K} \left| \mathbb{E}\Big[\underbrace{\big(Y^{(e)} - \beta_{\mathcal{S}(\beta)}^{\top} X_{\mathcal{S}(\beta)}^{(e)}\big)}_{\varepsilon^{(\mathbf{e})}} X_j^{(e)}\Big] \right|^2$$

Focused linear invariance regularizer

Population-level penalty:

★delete endogenous variables

$$J(\beta) = \sum_{j \in \mathcal{S}(\beta)} \sum_{e=1}^{K} \left| \mathbb{E}\big[\underbrace{\big(Y^{(e)} - \beta_{\mathcal{S}(\beta)}^{\top} X_{\mathcal{S}(\beta)}^{(e)} \big)}_{\varepsilon^{(\mathbf{e})}} X_j^{(e)} \big] \right|^2$$

★ If S is selected, minimizing $J(\beta)$ encourages $X_j^{(e)}$ and $\varepsilon^{(e)}$ uncorrelated across for all $j \in S$ and all environments.

A multi-environment version of linear least squares

★ **Population-level EILLS**: environment-invariant linear least-squares

$$Q(\beta;\gamma) = R(\beta) + \gamma J(\beta)$$

★ **EILLS estimator** $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma)$. $(\mathbb{E} \rightsquigarrow \widehat{\mathbb{E}})$

★ **Regularized EILLS estimator**: $\widehat{\beta}_L = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma) + \lambda\|\beta\|_0$.

A multi-environment version of linear least squares

★ **Population-level EILLS**: environment-invariant linear least-squares

$$Q(\beta;\gamma) = R(\beta) + \gamma J(\beta) = \sum_{e=1}^{K} \mathbb{E}[|Y^{(e)} - \beta^\top X^{(e)}|^2]$$

$$+\gamma \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} \times \sum_{e=1}^{K} \left|\mathbb{E}[(Y^{(e)} - \beta^\top X^{(e)})X_j^{(e)}]\right|^2$$

delete endogenous variables

★ **EILLS estimator** $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma)$. ($\mathbb{E} \rightsquigarrow \widehat{\mathbb{E}}$)

★ **Regularized EILLS estimator**: $\widehat{\beta}_L = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma) + \lambda\|\beta\|_0$.

A multi-environment version of linear least squares

★ **Population-level EILLS**: environment-invariant linear least-squares

$$Q(\beta;\gamma) = R(\beta) + \gamma J(\beta) = \sum_{e=1}^{K} \mathbb{E}[|Y^{(e)} - \beta^\top X^{(e)}|^2]$$

$$+\gamma \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} \times \sum_{e=1}^{K} \left|\mathbb{E}[(Y^{(e)} - \beta^\top X^{(e)})X_j^{(e)}]\right|^2$$

delete endogenous variables

★ **EILLS estimator** $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma)$. ($\mathbb{E} \rightsquigarrow \widehat{\mathbb{E}}$)

★ **Regularized EILLS estimator**: $\widehat{\beta}_L = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma) + \lambda\|\beta\|_0$.

A multi-environment version of linear least squares

★ **Population-level EILLS**: environment-invariant linear least-squares

$$Q(\beta;\gamma) = R(\beta) + \gamma J(\beta) = \sum_{e=1}^{K} \mathbb{E}[|Y^{(e)} - \beta^\top X^{(e)}|^2]$$

$$+\gamma \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}} \times \sum_{e=1}^{K} \left|\mathbb{E}[(Y^{(e)} - \beta^\top X^{(e)}) X_j^{(e)}]\right|^2$$

delete endogenous variables

★ **EILLS estimator** $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma)$. $(\mathbb{E} \rightsquigarrow \widehat{\mathbb{E}})$

★ **Regularized EILLS estimator**: $\widehat{\beta}_L = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma) + \lambda \|\beta\|_0.$

delete exog. var.

Two type of Spurious Variables

$\boxed{\textbf{Exogenous}} \ \boxed{\Longrightarrow}$ variance

Two type of Spurious Variables

$\boxed{\textbf{Exogenous}}$ $\boxed{\Longrightarrow}$ variance

$\boxed{\textbf{Endogenous}}$ $\boxed{\Longrightarrow}$ biases + incon.

Two type of Spurious Variables

$\boxed{\textbf{Exogenous}}$ $\boxed{\Longrightarrow}$ variance

reduced by $\ell_0(\beta)$ or $\|\beta\|_1$ or SCAD

$\boxed{\textbf{Endogenous}}$ $\boxed{\Longrightarrow}$ biases + incon.

Two type of Spurious Variables

$\boxed{\textbf{Exogenous}}$ $\boxed{\Longrightarrow}$ variance

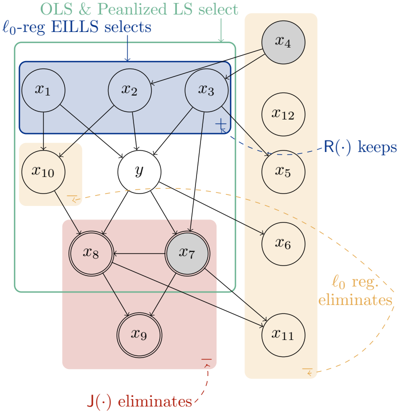
reduced by $\ell_0(\beta)$ or $\|\beta\|_1$ or SCAD

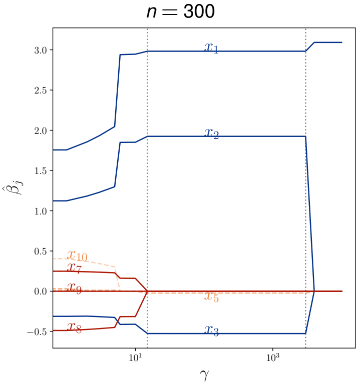
$\boxed{\textbf{Endogenous}}$ $\boxed{\Longrightarrow}$ biases + incon.

reduced by $J(\beta)$

How is S^* selected in SCM?

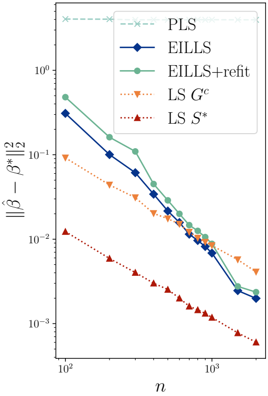
- $p = 12$, $S^* = \{1,2,3\}$, $\mathcal{G} = \{7,8,9\}$ (double circled).
- $e = 1$ observational env
- $e = 2$ interventional env: intervene on x_4, x_7 (shaded)



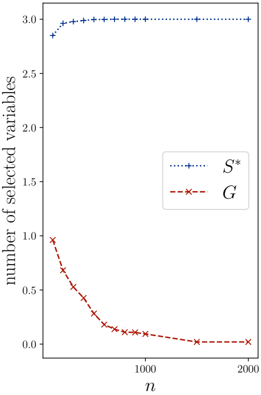


Simulation Results ($\gamma = 20$)

ℓ_2 error rate

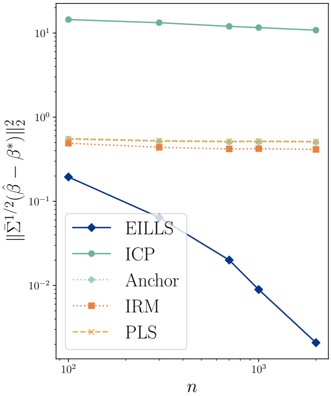


Variable selection



$N_{sim} = 500$

Comparisons



Non-asymptotic Result for EILLS

EILLS estimator: $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma)$.

Theorem 2. Under Cond 1-3 & IDF, if $\gamma \geq C\gamma^*$ and $p\gamma = o(n)$, then

(1) **Sure screening**: $S^* \subseteq \mathrm{supp}(\widehat{\beta}_Q) \subseteq \mathcal{G}^c$ holds w.h.p. for large n.

(2) **ℓ_2-error**. With high probability,

$$\|\widehat{\beta}_Q - \beta^*\|_2 \leq C\gamma\left\{\sqrt{\frac{|\mathcal{G}^c|}{n \cdot K}} + \frac{|\mathcal{G}^c|}{n}\right\};$$

Endogenous spurious: $\boxed{\mathcal{G} = \left\{j : \sum_{e=1}^K \mathbb{E}[X_j^{(e)}\varepsilon^{(e)}] \neq 0\right\}.}$

Non-asymptotic Result for EILLS

EILLS estimator: $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma)$.

Theorem 2. Under Cond 1-3 & IDF, if $\gamma \geq C\gamma^*$ and $p\gamma = o(n)$, then

(1) **Sure screening**: $S^* \subseteq \mathrm{supp}(\widehat{\beta}_Q) \subseteq \mathcal{G}^c$ holds w.h.p. for large n.

(2) **ℓ_2-error**. With high probability,

$$\|\widehat{\beta}_Q - \beta^*\|_2 \leq C\gamma \left\{ \sqrt{\frac{|\mathcal{G}^c|}{n \cdot K}} + \frac{|\mathcal{G}^c|}{n} \right\};$$

Selection consistency?

Non-asymptotic Result for EILLS

EILLS estimator: $\widehat{\beta}_Q = \mathrm{argmin}_\beta \widehat{Q}(\beta; \gamma)$.

Theorem 2. Under Cond 1-3 & IDF, if $\gamma \geq C\gamma^*$ and $p\gamma = o(n)$, then

(1) **Sure screening**: $S^* \subseteq \mathrm{supp}(\widehat{\beta}_Q) \subseteq \mathcal{G}^c$ holds w.h.p. for large n.

(2) **ℓ_2-error**. With high probability,

$$\|\widehat{\beta}_Q - \beta^*\|_2 \leq C\gamma \left\{ \sqrt{\frac{|\mathcal{G}^c|}{n \cdot K}} + \frac{|\mathcal{G}^c|}{n} \right\};$$

Endogenous Spurious $\times$ by $J(\beta)$

Non-asymptotic Result for EILLS

EILLS estimator: $\widehat{\beta}_Q = \text{argmin}_\beta \widehat{Q}(\beta; \gamma)$.

Theorem 2. Under Cond 1-3 & IDF, if $\gamma \geq C\gamma^*$ and $p\gamma = o(n)$, then

(1) **Sure screening**: $S^* \subseteq \text{supp}(\widehat{\beta}_Q) \subseteq \mathcal{G}^c$ holds w.h.p. for large n.

(2) **ℓ_2-error**. With high probability,

$$\|\widehat{\beta}_Q - \beta^*\|_2 \leq C\gamma \left\{ \sqrt{\frac{|\mathcal{G}^c|}{n \cdot K}} + \frac{|\mathcal{G}^c|}{n} \right\};$$

Endogenous Spurious ✗ by $J(\beta)$

Exogenous Spurious ✗ by $\ell_0(\beta)$

Variable Selection Consistency in High-dims

Regularized EILLS: $$\widehat{\beta}_L = \mathrm{argmin}_\beta \widehat{Q}(\beta;\gamma) + \lambda\|\beta\|_0.$$

Theorem 3. Under Conditions 1-3 & IDF, if $\gamma \geq C\gamma^*$, for sufficiently large n and proper choice of λ, we have

$$\mathbb{P}[\mathrm{supp}(\widehat{\beta}_L) = S^*] \geq 1 - p^{-10}.$$

★When $|S^\star| + \gamma = O(1)$, choose $\{K^{-1} + \sqrt{\frac{\log p}{n}}\}\frac{\log p}{n} \ll \lambda \ll \beta_{\min}^2$.

Neural Causal Learning

★ Gu, Y., Fang, C., Buelhmann, P., and Fan, J. (2024). Causality Pursuit from Heterogeneous Environments via Neural Adversarial Invariance Learning. *arxiv.org*

Nonparametric Causality Pursuit

★ Collect n data from K **heterogeneous** environment with dist $\mu^{(e)}$. For $e \in [K]$,

$$\boxed{Y^{(e)} = m^\star(X_{S^\star}^{(e)}) + \varepsilon^{(e)} \quad \text{with} \quad \mathbb{E}[\varepsilon^{(e)} | X_{S^\star}^{(e)}] = 0}$$

—$S^\star$ **unknown** variable set; —$m^\star$ **invariant** assoc.

♦**Much** weaker that standard reg: $\mathbb{E}[\varepsilon | X] = 0$.

★ **Goal**: estimate $S^\star$ and $m^\star$ using $n \cdot K$ data.

Focused Adversarial Invariance Regularizer (FAIR)

★ **Endogeneity (FAIR) Penalty**: ♣delete endogenous variables

$$\max_{\mathbf{f}\in\mathbf{S_g}} \left\{ \sum_{\mathbf{e}\in[\mathbf{K}]} \mathbb{E}_{\mu^{(\mathbf{e})}} \left[\{\mathbf{Y} - \mathbf{g}(\mathbf{X})\} \mathbf{f_e}(\mathbf{X}) \right] \right\} \text{ with } \mathbb{E}_{\mu^{(e)}} f_e^2(X) = 1.$$

—S_g is the support of function g

★ When $\mathrm{supp}(g) = S$, maximizing all $f_e(X_S)$ gives

$$\frac{1}{2} \sum_{e\in[K]} \mathbb{E}_{\mu^{(e)}} \left[\left| \mathbb{E}[Y|X_S] - g(X_S) \right|^2 \right].$$

Focused Adversarial Invariance Regularizer (FAIR)

★ **Endogeneity (FAIR) Penalty**: ♣delete endogenous variables

$$\max_{\mathbf{f}\in \mathbf{S_g}} \left\{ \sum_{\mathbf{e}\in[\mathbf{K}]} \mathbb{E}_{\mu^{(\mathbf{e})}} \left[\{\mathbf{Y} - \mathbf{g}(\mathbf{X})\}\mathbf{f_e}(\mathbf{X}) - \lambda f_e^2(X) \right] \right\}.$$

—S_g is the support of function g

★ When $\mathrm{supp}(g) = S$, maximizing all $f_e(X_S)$ gives

$$\frac{1}{2} \sum_{e\in[K]} \mathbb{E}_{\mu^{(e)}} \left[|\mathbb{E}[Y|X_S] - g(X_S)|^2 \right].$$

Focused Adversarial Invariance Regularizer (FAIR)

★ **Endogeneity (FAIR) Penalty**:

♣delete endogenous variables

$$J(g) = \max_{\mathbf{f}\in \mathbf{S_g}} \left\{ \sum_{\mathbf{e}\in[\mathbf{K}]} \mathbb{E}_{\mu^{(\mathbf{e})}} \left[\{\mathbf{Y} - \mathbf{g}(\mathbf{X})\} \mathbf{f_e}(\mathbf{X}) - \frac{1}{2} f_e^2(X) \right] \right\}.$$

—S_g is the support of function g

★ When $\mathrm{supp}(g) = S$, maximizing all $f_e(X_S)$ gives

$$\frac{1}{2} \sum_{e\in[K]} \mathbb{E}_{\mu^{(e)}} \left[|\mathbb{E}[Y|X_S] - g(X_S)|^2 \right].$$

Focused Adversarial Invariance Regularizer (FAIR)

★ **Endogeneity (FAIR) Penalty**: ♣delete endogenous variables

$$J(g) = \max_{\mathbf{f} \in \mathbf{S_g}} \left\{ \sum_{\mathbf{e} \in [\mathbf{K}]} \mathbb{E}_{\mu^{(\mathbf{e})}} \left[\{\mathbf{Y} - \mathbf{g}(\mathbf{X})\} \mathbf{f_e}(\mathbf{X}) - \frac{1}{2} f_e^2(X) \right] \right\}.$$

—S_g is the support of function g

★ When $\operatorname{supp}(g) = S$, maximizing all $f_e(X_S)$ gives

$$\frac{1}{2} \sum_{e \in [K]} \mathbb{E}_{\mu^{(e)}} \left[\left| \mathbb{E}[Y | X_S] - g(X_S) \right|^2 \right].$$

FAIR Estimation Method

■ Predictor class $\mathcal{G}$, Discriminator class $\mathcal{F}$.

★ Population-level Objective Function:

$$Q(g, f; \gamma) = \sum_{e \in [K]} \mathbb{E}_{\mu^{(e)}} [\ell(g(X), Y)] + \gamma J(g)$$

★ Empirical FAIR Estimator: $\mathbb{E} \rightsquigarrow \widehat{\mathbb{E}}$

$$\widehat{g} \in \operatorname{argmin}_{g \in \mathcal{G}} \max_{f \in \mathcal{F}_{S_g}} \widehat{Q}(g, f; \gamma)$$

FAIR Estimation Method

■ Predictor class $\mathcal{G}$, Discriminator class $\mathcal{F}$.

★ Population-level Objective Function:

$$Q(g, f; \gamma) = \sum_{e \in [K]} \mathbb{E}_{\mu^{(e)}} [\ell(g(X), Y)] + \gamma J(g)$$

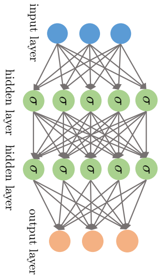
★ Empirical FAIR Estimator: $\mathbb{E} \rightsquigarrow \widehat{\mathbb{E}}$

$$\widehat{g} \in \operatorname{argmin}_{g \in \mathcal{G}} \max_{f \in \mathcal{F}_{S_g}} \widehat{Q}(g, f; \gamma)$$

Causal Adversarial Networks

FAIR-NN:

- ★ $\mathcal{G}$: ReLU network with width N and depth L.
- ★ $\mathcal{F}$: ReLU network with width $2N$ and depth $L+2$.



Identifiability: IDF-A

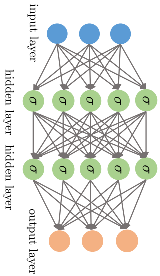
★ $m^{(e,S)}(x) = \mathbb{E}[Y^{(e)}|X_S = x_S]$

$$\forall S \text{ if } \bar{m}^{(S\cup S^\star)} \neq m^\star \Longrightarrow \exists e, e' \in [K], \text{ s.t. } m^{(e,S)} \neq m^{(e',S)}$$

Causal Adversarial Networks

FAIR-NN:

- ★ $\mathcal{G}$: ReLU network with width N and depth L.
- ★ $\mathcal{F}$: ReLU network with width $2N$ and depth $L+2$.



Identifiability: `IDF-A`

★$m^{(e,S)}(x) = \mathbb{E}[Y^{(e)}|X_S = x_S]$

$$\forall S \text{ if } \bar{m}^{(S \cup S^\star)} \neq m^\star \Longrightarrow \exists e, e' \in [K], \text{ s.t. } m^{(e,S)} \neq m^{(e',S)}$$

Properties for FAIR-NN

$$\gamma^\star = \sup_{S: m^\star \neq \bar{m}^{(S\cup S^\star)}} \frac{\|\bar{m}^{(S\cup S^\star)} - m^\star\|_2^2}{\frac{1}{|\mathcal{E}|}\|m^{(e,S)} - \bar{m}^{(S)}\|_{2,e}^2}$$

Bias of LS w/ all data

Variance of biases

Theorem 4. (Oracle-type of Inequality) END

Under Condiions IDF-A, if $\gamma \geq 8\gamma^*$, for large enough n,

$$\|\widehat{g} - m^\star\|_2 \leq \tilde{C}\left\{\inf_{g\in\mathcal{G}_{S^\star}} \|g - m^\star\|_2 + \frac{NL}{\sqrt{n}}\right\}.$$

1. Rates depends on approx. errors of m^*

Adaptive Learning

2. For HCM $m^* = f_1 \circ \cdots \circ f_q$, rate is $n^{-\frac{\alpha^*}{2\alpha^*+1}}$, with $\alpha^* = \min(\beta_j/d_j)$.

   e.g. $m^* = f_1(x_1) + \cdots + f_p(x_p)$ $\quad m^* = f_1(x_1, f_2(x_2, x_3)) + f_4(x_2, x_9)$

Properties for FAIR-NN

$$\gamma^\star = \sup_{S: m^\star \neq \bar{m}^{(S \cup S^\star)}} \frac{\|\bar{m}^{(S \cup S^\star)} - m^\star\|_2^2}{\frac{1}{|\mathcal{E}|}\|m^{(e,S)} - \bar{m}^{(S)}\|_{2,e}^2}$$

Bias of LS w/ all data

Variance of biases

Theorem 4. (Oracle-type of Inequality) END

Under Condiions `IDF-A`, if $\gamma \geq 8\gamma^*$, for large enough n,

$$\|\widehat{g} - m^\star\|_2 \leq \tilde{C}\left\{\inf_{g \in \mathcal{G}_{S^\star}} \|g - m^\star\|_2 + \frac{NL}{\sqrt{n}}\right\}.$$

1. Rates depends on approx. errors of m^*

Adaptive Learning

2. For HCM $m^* = f_1 \circ \cdots \circ f_q$, rate is $n^{-\frac{\alpha^*}{2\alpha^*+1}}$, with $\alpha^* = \min(\beta_j/d_j)$.

   e.g. $m^* = f_1(x_1) + \cdots + f_p(x_p)$ $\qquad$ $m^* = f_1(x_1, f_2(x_2, x_3)) + f_4(x_2, x_9)$

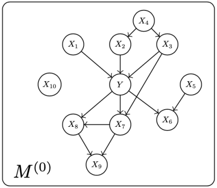
Causality under SCM

$S^\star$ **is direct causes under non-degenerate interventions**

Structural Causal Model with Intervention

SCM Model: For each env $e \in \mathcal{E}$, $(X^{(e)}, Y^{(e)}) = (Z_1^{(e)}, \ldots, Z_d^{(e)}, Z_{d+1}^{(e)})$

$$X_j^{(e)} \leftarrow f_j^{(e)}(Z_{\mathrm{pa}(j)}^{(e)}, U_j) \quad \forall j \in [d], \qquad Y^{(e)} \leftarrow f_{d+1}(X_{\mathrm{pa}(d+1)}^{(e)}, U_{d+1})$$



Intervention: Some X_I intervened: SCM $\tilde{M}$ of (X, Y, E) is $E \leftarrow \mathrm{Unif}([K])$

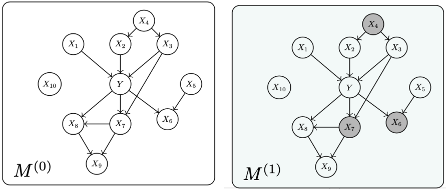
$$X_j \leftarrow \begin{cases} f_j(Z_{\mathrm{pa}(j)}, U_j, \mathbf{E}) & \forall j \in I \\ f_j(Z_{\mathrm{pa}(j)}, U_j) & \forall j \in [d] \setminus I \end{cases} \qquad Y \leftarrow f_{d+1}(X_{\mathrm{pa}(d+1)}, U_{d+1})$$

★DAG induced graph ★Unknown interventions, not on Y.

Structural Causal Model with Intervention

SCM Model: For each env $e \in \mathcal{E}$, $(X^{(e)}, Y^{(e)}) = (Z_1^{(e)}, \ldots, Z_d^{(e)}, Z_{d+1}^{(e)})$

$$X_j^{(e)} \leftarrow f_j^{(e)}(Z_{\mathrm{pa}(j)}^{(e)}, U_j) \quad \forall j \in [d], \qquad Y^{(e)} \leftarrow f_{d+1}(X_{\mathrm{pa}(d+1)}^{(e)}, U_{d+1})$$



$I = \{4,6,7\}$

Intervention: Some X_I intervened: SCM $\tilde{M}$ of (X, Y, E) is $E \leftarrow \mathrm{Unif}([K])$

$$X_j \leftarrow \begin{cases} f_j(Z_{\mathrm{pa}(j)}, U_j, \mathbf{E}) & \forall j \in I \\ f_j(Z_{\mathrm{pa}(j)}, U_j) & \forall j \in [d] \setminus I \end{cases} \qquad Y \leftarrow f_{d+1}(X_{\mathrm{pa}(d+1)}, U_{d+1})$$

★DAG induced graph ★Unknown interventions, not on Y.

Characterizing $S^\star$ under Intervention

Theorem 5. Existence of Maximum Invariant Set

Under nondegenerate interventions, Condition `IDF-A` holds with

$$S^\star = \mathtt{pa}(d+1) \cup A(I) \cup \bigcup_{j \in A(I)} (\mathtt{pa}(j) \setminus \{d+1\})$$

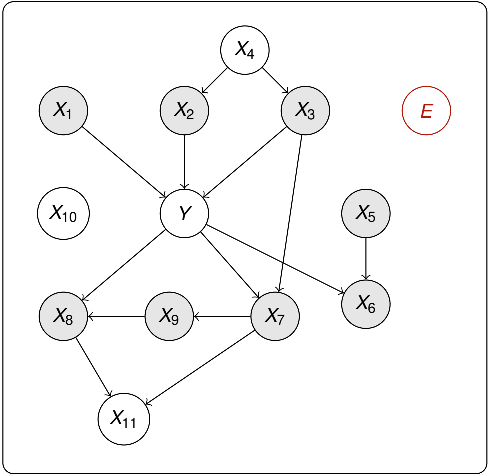
where $A(I) = \{j : j \in \mathtt{ch}(d+1), \mathtt{at}(j) \cap \mathtt{ch}(d+1) \cap I = \emptyset\}$

Invariant variables: ★parents of Y;

★uninterviewed children of Y;

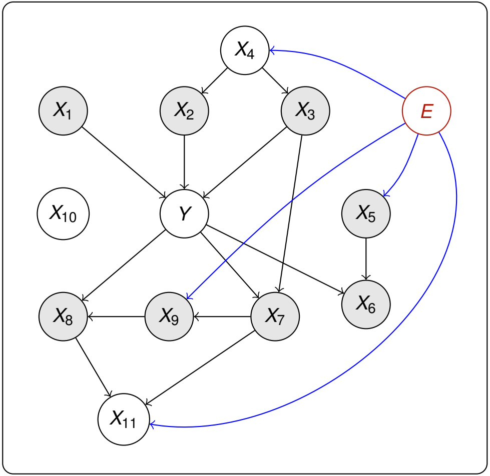
★parents of uninterviewed children of Y.

An Illustration: $k = 1$



$0 \leftrightarrow 0$, $S^\star = \{1, 2, 3, 5, 6, 7, 8, 9\}$

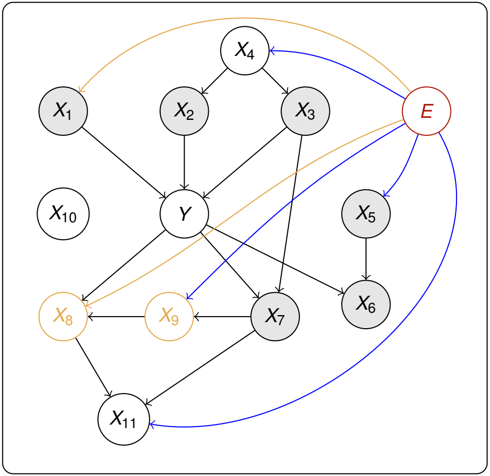
An Illustration: $k = 2$



$0 \leftrightarrow 0$, $S^\star = \{1, 2, 3, 5, 6, 7, 8, 9\}$

$0 \leftrightarrow \mathbf{1}$, $S^\star = \{1, 2, 3, 5, 6, 7, 8, 9\}$

An Illustration: $k = 3$

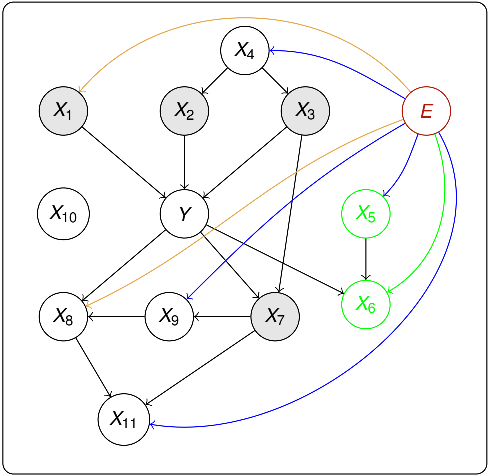


$0 \leftrightarrow 0$, $S^\star = \{1,2,3,5,6,7,8,9\}$

$0 \leftrightarrow \mathbf{1}$, $S^\star = \{1,2,3,5,6,7,8,9\}$

$0 \leftrightarrow 2$, $S^\star = \{1,2,3,5,6,7\}$

An Illustration: $k = 4$



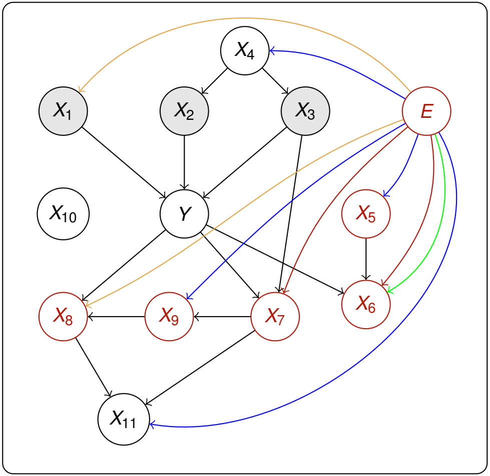
$0 \leftrightarrow 0,\ S^\star = \{1, 2, 3, 5, 6, 7, 8, 9\}$

$0 \leftrightarrow \mathbf{1},\ S^\star = \{1, 2, 3, 5, 6, 7, 8, 9\}$

$0 \leftrightarrow 2,\ S^\star = \{1, 2, 3, 5, 6, 7\}$

$0 \leftrightarrow \mathbf{3},\ S^\star = \{1, 2, 3, 7\}$

An Illustration: $k = 5$



$0 \leftrightarrow 0,\ S^\star = \{1,2,3,5,6,7,8,9\}$

$0 \leftrightarrow \mathbf{1},\ S^\star = \{1,2,3,5,6,7,8,9\}$

$0 \leftrightarrow 2,\ S^\star = \{1,2,3,5,6,7\}$

$0 \leftrightarrow \mathbf{3},\ S^\star = \{1,2,3,7\}$

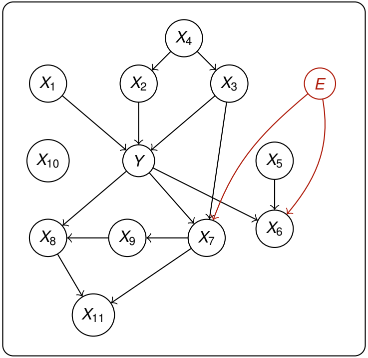
$0 \leftrightarrow 4,\ S^\star = \{1,2,3\}$

Exact Direct Causal Recovery

Proposition 1. Sufficient and Necessary Condition for Causal Discovery

When all root-children are intervened(★), $S^\star = \mathrm{pa}(d+1)$. The condition is also necessary, if Y does not have degenerate children.

★ $I \supseteq I^\star$, where $I^\star = \{j : j \in \mathrm{ch}(d+1), \mathrm{pa}(j) \cap \mathrm{ch}(d+1) = \emptyset\}$.

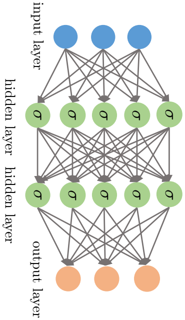


$0 \leftrightarrow 4$

$S^\star = \{1, 2, 3\}$

Implementation and Simulations

Challenge of Implementation



Parameterization: g_θ, f_{e,ϕ_e} with $e \in [K]$

Objective: $\widehat{g} \in \mathrm{argmin}_{g \in \mathcal{G}} \max_{\{f_e \in \mathcal{F}_{S_g}\}_{e \in [K]}} \widehat{\mathrm{Q}}(g, f^{[K]}; \gamma)$

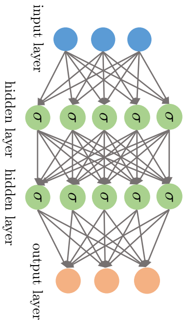
★ min-max optimization. $\rightsquigarrow$ gradient descent ascent

★ f has the same X-variables as g.

$$\min_{\theta, a \in \{0,1\}^d} \max_{\phi_1, \ldots, \phi_K} \mathcal{L}(g_\theta(a \odot x), \{f_{e,\phi_e}(a \odot x)\}_{e=1}^K)$$

$\rightsquigarrow$ **Enumerate all $a \in \{0,1\}^d$!**

Challenge of Implementation



Parameterization: g_θ, f_{e,ϕ_e} with $e \in [K]$

Objective: $\widehat{g} \in \mathrm{argmin}_{g \in \mathcal{G}} \max_{\{f_e \in \mathcal{F}_{S_g}\}_{e\in[K]}} \widehat{\mathrm{Q}}(g, f^{[K]}; \gamma)$

- ★ min-max optimization. $\rightsquigarrow$ gradient descent ascent
- ★ f has the same X-variables as g.

$$\min_{\theta, a \in \{0,1\}^d} \max_{\phi_1,\ldots,\phi_k} \mathcal{L}(g_\theta(a \odot x), \{f_{e,\phi_e}(a \odot x)\}_{e=1}^K)$$

$\rightsquigarrow$ **Enumerate all $a \in \{0,1\}^d$!**

Gradient Method with Gumbel Approximation

Equivalence Problem: $\sigma(u) = 1/(1+e^{-u})$

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{A\sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A \odot x), \{f_{e,\phi_e}(A \odot x)\}_{e=1}^{K})$$

Gradient Method with Gumbel Approximation

$$\min_{\theta, w} \max_{\phi_1, \ldots, \phi_k} \mathbb{E}_{A \sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A \odot x), \{f_{e,\phi_e}(A \odot x)\}_{e=1}^{K})$$

Gumbel Approx: $\mathrm{Bern}(\sigma(w)) = I(U - \sigma(w) < 0)$

Gradient Method with Gumbel Approximation

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{A \sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A \odot x), \{f_{e,\phi_e}(A \odot x)\}_{e=1}^{K})$$

Gumbel Approx: $\mathrm{Bern}(\sigma(w)) = I(U - \sigma(w) < 0) = I(\mathrm{logit}(U) - w < 0)$

Gradient Method with Gumbel Approximation

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{A \sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A \odot x), \{f_{e,\phi_e}(A \odot x)\}_{e=1}^{K})$$

Gumbel Approx: $\mathrm{Bern}(\sigma(w)) = I(U - \sigma(w) < 0) \approx \frac{1}{1+\exp((\mathrm{logit}(U) - w)/\tau)}$, as $\tau \to 0$

Gradient Method with Gumbel Approximation

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{A\sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A\odot x), \{f_{e,\phi_e}(A\odot x)\}_{e=1}^K)$$

Gumbel Approx: $\mathrm{Bern}(\sigma(w)) = I(U - \sigma(w) < 0) \approx \frac{1}{1+\exp((V_2 - V_1 - w)/\tau)} \equiv B_\tau(V, w)$ as $\tau \to 0$

Gradient Method with Gumbel Approximation

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{A\sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A \odot x), \{f_{e,\phi_e}(A \odot x)\}_{e=1}^K)$$

Gumbel Approx: $\mathrm{Bern}(\sigma(w)) = I(U - \sigma(w) < 0) \approx \frac{1}{1+\exp((V_2 - V_1 - w)/\tau)} \equiv B_\tau(V, w)$ as $\tau \to 0$

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{V\sim \mathrm{Gum}} \mathcal{L}(g_\theta(B_\tau(V,w) \odot x), \{f_{e,\phi_e}(B_\tau(V,w) \odot x)\}_{e=1}^K)$$

Gradient Method with Gumbel Approximation

$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{A\sim \mathrm{Bern}(\sigma(w))} \mathcal{L}(g_\theta(A \odot x), \{f_{e,\phi_e}(A \odot x)\}_{e=1}^K)$$

Gumbel Approx: $\mathrm{Bern}(\sigma(w)) = I(U - \sigma(w) < 0) \approx \frac{1}{1+\exp((V_2 - V_1 - w)/\tau)} \equiv B_\tau(V, w)$ as $\tau \to 0$

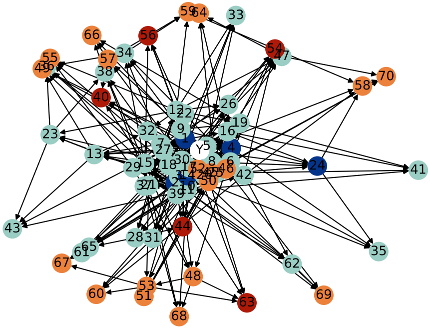
$$\min_{\theta, w} \max_{\phi_1,\ldots,\phi_k} \mathbb{E}_{V\sim \mathrm{Gum}} \mathcal{L}(g_\theta(B_\tau(V, w) \odot x), \{f_{e,\phi_e}(B_\tau(V, w) \odot x)\}_{e=1}^K)$$

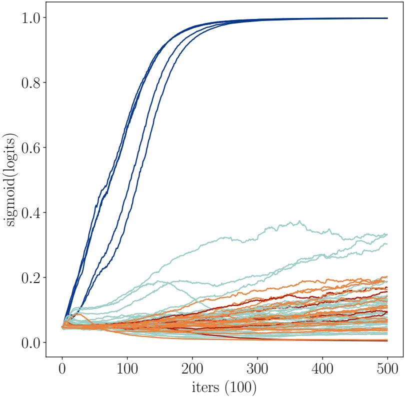
Algorithm:

1. Sample V, batch of $(X^{(e)}, Y^{(e)})$ from each environment.
2. Gradient ascent update for $\phi_1, \ldots, \phi_k$.
3. Gradient descent update for θ, w, with decreasing temperature τ.

Linear Model, $\mathcal{G} = \mathcal{F}$ linear

- $k = 2$, $d = 70$, brute force search is impossible
- Random generated SCM sharing same cause-effect relationship.
- All X are intervened (randomly).

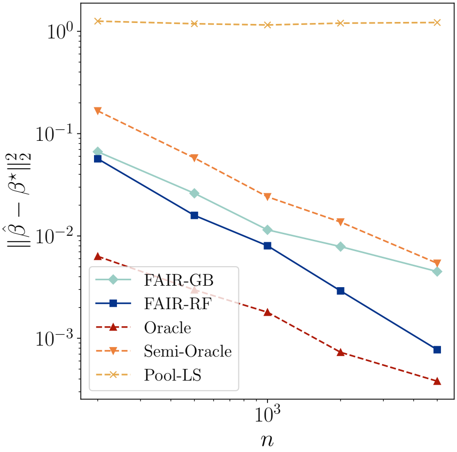


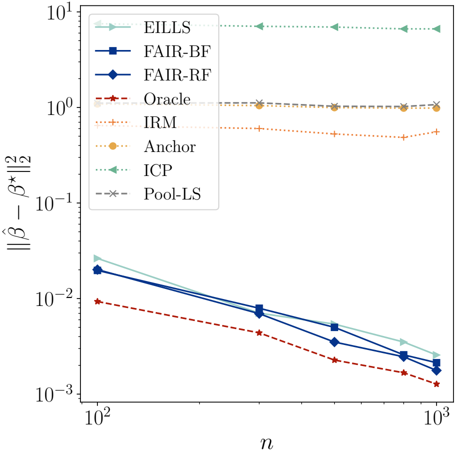


Relations with Y: blue = parent, red = child, orange = offspring, lightblue = other

Performance of FAIR-Linear

- FAIR-GB: implementation using Gumbel approximation.
- FAIR-RF: refitting after running FAIR-GB.





★log-log plot of med(MSE) based on $N_{sim} = 50$ for •(a) $p = 70$, $n \in \{200, 500, 1000, 2000, 5000\}$ and •$p = 15$ and $n \in \{100, 200, 500, 800, 1000\}$

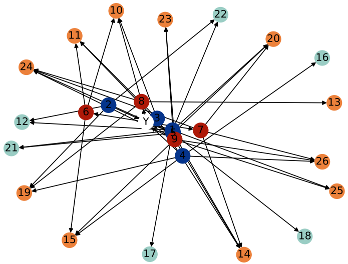
Simulations for FAIR-NN

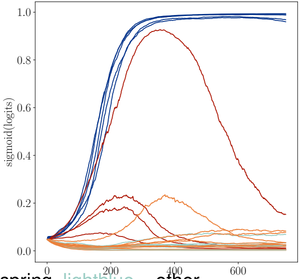
$d = 26$, $k = 2$

$$X_i^{(e)} \leftarrow \begin{cases} \varepsilon_i^{(e)} & i \le 5 \\ f_{i,0}^{(e)}(Y^{(e)}) + \varepsilon_i^{(e)} & 6 \le i \le 9 \\ \sum_{j \in \mathrm{pa}(i) \subseteq [8]} f_{i,j}^{(e)}(X_j^{(e)}) + \varepsilon_i^{(e)} & 10 \le i \le 26 \end{cases}$$

$$Y^{(e)} \leftarrow m_k^\star(X_1^{(e)}, \ldots, X_5^{(e)}) + \varepsilon_0,$$

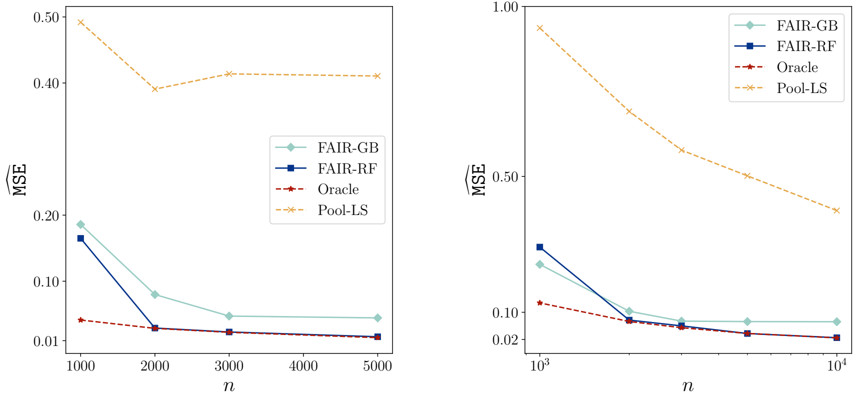
$m_1^\star(x)$ additive, $m_2^\star(x) = x_1 x_2^3 + \log(1 + e^{\tanh(x_3)} + e^{x_4}) + \sin(x_5)$ HCM.





Relations with Y: blue = parent, red = child, orange = offspring, lightblue = other

Performance of FAIR-NN



★MSE over $N_{sim} = 50$ over 30K x-values. •(a) additive $m_1^\star$ and $n \in \{1000, 2000, 3000, 5000\}$

•(b) $m_2^\star$ and $n \in \{1000, 2000, 3000, 5000, 10000\}$.

Application I: Transfer Learning

Waterbird Classification

- $Y = 1$ (**water bird**) and $Y = 0$ (**land bird**)
- $X \in \mathbb{R}^{500}$ extracted from ResNet pre-trained on ImageNet.

Data

★ Training data with spurious background (n=50k).

- $\mathcal{D}_1$: 95% **water birds** on **water**, 90% **land birds** on **land**.
- $\mathcal{D}_2$: 75% **water birds** on **water**, 70% **land birds** on **land**.

★ Test data with reverse spurious background (n=30k).

- $\mathcal{D}_3$: 98% **water birds** on **land**, 98% **land birds** on **water**.

—Build a linear classifier using $(\mathcal{D}_1, \mathcal{D}_2)$.

Application I: Transfer Learning

Waterbird Classification

- $Y = 1$ (**water bird**) and $Y = 0$ (**land bird**)
- $X \in \mathbb{R}^{500}$ extracted from ResNet pre-trained on ImageNet.

Data

★Training data with spurious background (n=50k).
- $\mathcal{D}_1$: 95% **water birds** on **water**, 90% **land birds** on **land**.
- $\mathcal{D}_2$: 75% **water birds** on **water**, 70% **land birds** on **land**.

★Test data with reverse spurious background (n=30k).
- $\mathcal{D}_3$: 98% **water birds** on **land**, 98% **land birds** on **water**.

—Build a linear classifier using $(\mathcal{D}_1, \mathcal{D}_2)$.

Application I: Transfer Learning

Waterbird Classification

- $Y = 1$ (**water bird**) and $Y = 0$ (**land bird**)
- $X \in \mathbb{R}^{500}$ extracted from ResNet pre-trained on ImageNet.

Data

★ Training data with spurious background (n=50k).

- $\mathcal{D}_1$: 95% **water birds** on **water**, 90% **land birds** on **land**.
- $\mathcal{D}_2$: 75% **water birds** on **water**, 70% **land birds** on **land**.

★ Test data with reverse spurious background (n=30k).

- $\mathcal{D}_3$: 98% **water birds** on **land**, 98% **land birds** on **water**.

—Build a linear classifier using $(\mathcal{D}_1, \mathcal{D}_2)$.

Application I: Transfer Learning

Waterbird Classification

- $Y = 1$ (**water bird**) and $Y = 0$ (**land bird**)
- $X \in \mathbb{R}^{500}$ extracted from ResNet pre-trained on ImageNet.

Data

★Training data with spurious background (n=50k).

- $\mathcal{D}_1$: 95% **water birds** on **water**, 90% **land birds** on **land**.
- $\mathcal{D}_2$: 75% **water birds** on **water**, 70% **land birds** on **land**.

★Test data with reverse spurious background (n=30k).

- $\mathcal{D}_3$: 98% **water birds** on **land**, 98% **land birds** on **water**.

—Build a linear classifier using $(\mathcal{D}_1, \mathcal{D}_2)$.

Bias Reduction from Two Biased Samples

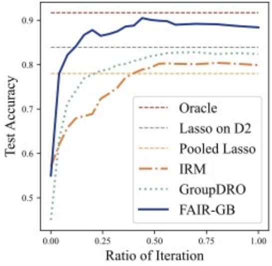
Methods and Results:

- ★ **FAIR-GB** FAIR estimator with linear $(\mathcal{G}, \mathcal{F})$, cross-entropy loss and Gumbel appox.
- ★ **PooledLasso** on $\mathcal{D}_1 \cup \mathcal{D}_2$; **Lasso on D2** Lasso on $\mathcal{D}_2$.
- ★ **Oracle**: Lasso on $\mathcal{D}_4$ where label/background independent.
- ★ **IRM** invariant risk minimization; **GroupDRO** group distributionally robust optimization.

Methods and Results:

- ★ **FAIR-GB** FAIR estimator with linear $(\mathcal{G}, \mathcal{F})$, cross-entropy loss and Gumbel appox.
- ★ **PooledLasso** on $\mathcal{D}_1 \cup \mathcal{D}_2$; **Lasso on D2** Lasso on $\mathcal{D}_2$.
- ★ **Oracle**: Lasso on $\mathcal{D}_4$ where label/background independent.
- ★ **IRM** invariant risk minimization; **GroupDRO** group distributionally robust optimization.

Method	Test Accuracy
Oracle	91.06 ± 0.24 %
Lasso on D2	84.57 ± 0.71 %
Pooled Lasso	79.08 ± 0.54 %
IRM	80.32 ± 0.67 %
GroupDRO	82.75 ± 1.10 %
FAIR-GB	89.56 ± 0.53 %

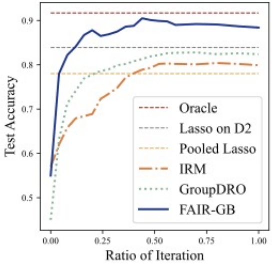


Bias Reduction from Two Biased Samples

Methods and Results:

- ★ **FAIR-GB** FAIR estimator with linear $(\mathcal{G}, \mathcal{F})$, cross-entropy loss and Gumbel appox.
- ★ **PooledLasso** on $\mathcal{D}_1 \cup \mathcal{D}_2$; **Lasso on D2** Lasso on $\mathcal{D}_2$.
- ★ **Oracle**: Lasso on $\mathcal{D}_4$ where label/background independent.
- ★ **IRM** invariant risk minimization; **GroupDRO** group distributionally robust optimization.

Method	Test Accuracy
Oracle	91.06 ± 0.24 %
Lasso on D2	84.57 ± 0.71 %
Pooled Lasso	79.08 ± 0.54 %
IRM	80.32 ± 0.67 %
GroupDRO	82.75 ± 1.10 %
FAIR-GB	89.56 ± 0.53 %

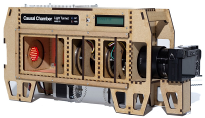


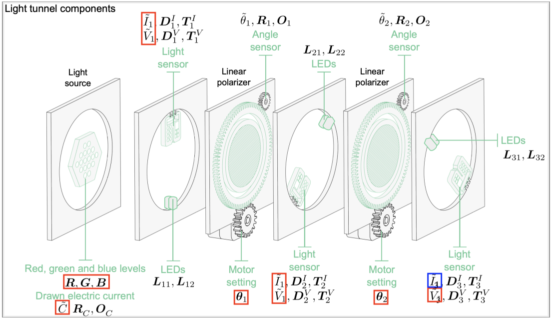
★**FAIR can correct bias from two biased samples!**

Application II: Causal Discovery in Physical Systems

Light Tunnel Device : **known causality** from physics — Gamella et al (2024)

$\rightsquigarrow$ **a testbed for causal discovery methods**





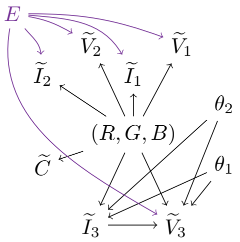
Task Find cause of $Y = \tilde{I}_3$ from covariates $X = (R, G, B, \theta_1, \theta_2, \tilde{V}_1, \tilde{V}_2, \tilde{V}_3, \tilde{I}_1, \tilde{I}_2, \tilde{C})$

Dataset ★$\mathcal{D}_0$: obs env (size 10000) ★D_1: weak intervene env (size 3000) on $(\tilde{V}_j)_{j=1}^3, (\tilde{I}_j)_{j=1}^2$

Data and Ground Truth

Data $\breve{D}_1, \breve{D}_2$ sub-sample of D_1, D_2 with equal size n.

Augmented SCM graph



★Direct Causes $S^{\star} = (R, G, B, \theta_1, \theta_2)$.

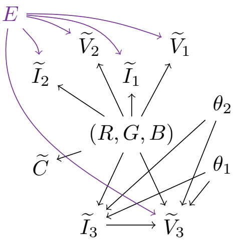
★Challenges

- weak & nonlinear signal $\tilde{I}_3 \propto \cos^2(\theta_1 - \theta_2)$.
- strong spurious association $\tilde{I}_3 \leftrightarrow \tilde{V}_3$.
- strong explained R^2 for $\tilde{I}_2, \tilde{I}_1$ (≥ 0.9).

Data and Ground Truth

Data $\breve{D}_1, \breve{D}_2$ sub-sample of D_1, D_2 with equal size n.

Augmented SCM graph



★Direct Causes $S^\star = (R, G, B, \theta_1, \theta_2)$.

★Challenges

- weak & nonlinear signal $\tilde{I}_3 \propto \cos^2(\theta_1 - \theta_2)$.
- strong spurious association $\tilde{I}_3 \leftrightarrow \tilde{V}_3$.
- strong explained R^2 for $\tilde{I}_2, \tilde{I}_1$ (≥ 0.9).

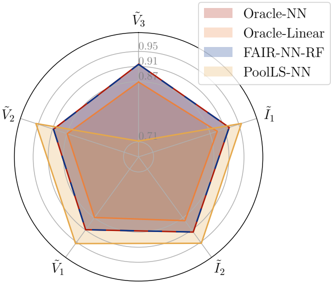
Prediction Performance Evaluation

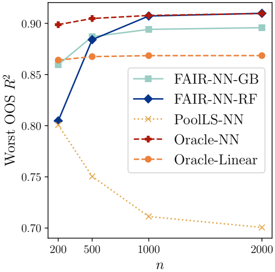
Methods

- ★ **FAIR-NN-GB**: Gumbel implented FAIR-NN, **FAIR-NN-RF** refitted estimator
- ★ **Oracle-*M***: Regress Y on $X_{S^\star}$ using $M \in \{\mathrm{Linear}, \mathrm{NN}\}$.
- ★ **PoolLS-NN**: Regress Y on $X_{S^\star}$ using all the data and NN.

Evaluate the Dependency on Variables Other Than $X_{S^\star}$

★Out-of-sample(OOS)-R^2 on $\mathcal{D}_{3,Z}$ with $Z \in \{\tilde{V}_j\}_{j=1}^3 \cup \{\tilde{I}_j\}_{j=1}^2$ where Z is **strongly** intervened.





n=1000 ♦Remove strong spurious var $\tilde{V}_3$ (otherwise $R^2 \downarrow 0.2$) ♦Detect weak signals (θ_1, θ_2): $R^2 \uparrow 0.04$ as Linear $\to$ NN.

$\rightsquigarrow$ **Near oracle performance.**

Attaining Variable Selection Consistency

Methods

- ★ **FAIR-M**: Gumbel implented FAIR-M method $M \in \{\mathrm{Linear}, \mathrm{NN}\}$, $\widehat{S} = \{j : \sigma(w_j) > 0.9\}$.
- ★ **ForestVarSel**: Select by importance measure using RandomForest
- ★ **NonlinearICP**: Previous invariance learning estimators.

Results (blue=parent, red=child, orange=neither ancestor nor descendants.)

		R	G	B	θ_1	θ_2	$\tilde{V}_1$	$\tilde{V}_2$	$\tilde{V}_3$	$\tilde{I}_1$	$\tilde{I}_2$	$\tilde{C}$
$n = 1000$	FAIR-Linear	1.0	1.0	1.0	0.09	0.06	0.0	0.0	0.0	0.0	0.0	0.01
	FAIR-NN	1.0	1.0	1.0	0.99	0.99	0.0	0.0	0.0	0.03	0.01	0.0
	ForestVarSel	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
	NonlinearICP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FAIR-NN	$n = 200$	1.0	0.99	0.74	0.61	0.65	0.04	0.02	0.29	0.04	0.04	0.01
	$n = 500$	1.0	1.0	0.95	0.86	0.89	0.03	0.03	0.03	0.01	0.02	0.0
	$n = 2000$	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

★Variable Selection Consistency ★NN detect nonlinear Malus's law $\tilde{I}_3 \propto \cos^2(\theta_1 - \theta_2)$.

Attaining Variable Selection Consistency

Methods

- ★ **FAIR-M**: Gumbel implented FAIR-M method $M \in \{\mathrm{Linear}, \mathrm{NN}\}$, $\widehat{S} = \{j : \sigma(w_j) > 0.9\}$.
- ★ **ForestVarSel**: Select by importance measure using RandomForest
- ★ **NonlinearICP**: Previous invariance learning estimators.

Results (blue=parent, red=child, orange=neither ancestor nor descendants.)

		R	G	B	θ_1	θ_2	$\tilde{V}_1$	$\tilde{V}_2$	$\tilde{V}_3$	$\tilde{I}_1$	$\tilde{I}_2$	$\tilde{C}$
$n = 1000$	FAIR-Linear	1.0	1.0	1.0	0.09	0.06	0.0	0.0	0.0	0.0	0.0	0.01
	FAIR-NN	1.0	1.0	1.0	0.99	0.99	0.0	0.0	0.0	0.03	0.01	0.0
	ForestVarSel	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
	NonlinearICP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FAIR-NN	$n = 200$	1.0	0.99	0.74	0.61	0.65	0.04	0.02	0.29	0.04	0.04	0.01
	$n = 500$	1.0	1.0	0.95	0.86	0.89	0.03	0.03	0.03	0.01	0.02	0.0
	$n = 2000$	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

★Variable Selection Consistency ★NN detect nonlinear Malus's law $\tilde{I}_3 \propto \cos^2(\theta_1 - \theta_2)$.

Summary

1. Introduce a FAIR-NN method to learn causal predictors w/o knowledge of

   ★cause-effect ★function structure

   - ♦ **Neural network** $\rightsquigarrow$ learn feature representation from data
   - ♦ **Invariance** $\rightsquigarrow$ distinguish causal/non-causal via **FAIR-penalty** J_0

2. Establish sample efficiency in different aspects.
   - Minimal identification condition.
   - Convergence rate depends on $m^\star$, adapt to low-dimension structures.
   - Regularization hyper-parameter minor impact.

3. Give an efficient implementation via Gumble Approx using SGD.

Summary

1. Introduce a FAIR-NN method to learn causal predictors w/o knowledge of

   ★cause-effect ★function structure

   - ♦ **Neural network** $\rightsquigarrow$ learn feature representation from data
   - ♦ **Invariance** $\rightsquigarrow$ distinguish causal/non-causal via **FAIR-penalty** J_0

2. Establish sample efficiency in different aspects.

   - Minimal identification condition.
   - Convergence rate depends on $m^\star$, adapt to low-dimension structures.
   - Regularization hyper-parameter minor impact.

3. Give an efficient implementation via Gumble Approx using SGD.

Summary

1. Introduce a FAIR-NN method to learn causal predictors w/o knowledge of

   ★cause-effect ★function structure

   - ♦ **Neural network** $\rightsquigarrow$ learn feature representation from data
   - ♦ **Invariance** $\rightsquigarrow$ distinguish causal/non-causal via **FAIR-penalty** J_0

2. Establish sample efficiency in different aspects.

   - Minimal identification condition.
   - Convergence rate depends on $m^\star$, adapt to low-dimension structures.
   - Regularization hyper-parameter minor impact.

3. Give an efficient implementation via Gumble Approx using SGD.

The End

Thank  *You*

★ Fan, J., Fang, C., Gu, Y., and Zhang, T. (2024+). Environment Invariant Linear Least Squares. AOS

★ Gu, Y., Fang, C., Buelhmann, P., and Fan, J. (2024). Causality Pursuit from Heterogeneous Environments via Neural Adversarial Invariance Learning. *arxiv.org*