Incorporating Causal Graphical Prior Knowledge into Predictive Modeling via Simple Data Augmentation

Takeshi Teshima[1,2], Masashi Sugiyama[2,1]

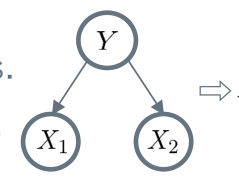
[1] The University of Tokyo [2] RIKEN

Overview

Causal Graphs (CGs) (Pearl, 2009)

Representation of our knowledge of data generating processes.

CGs imply conditional independence (CI) relations (Pearl, 2009) (Richardson, 2003).



$\Rightarrow X_1 \perp\!\!\!\perp X_2 \mid Y$

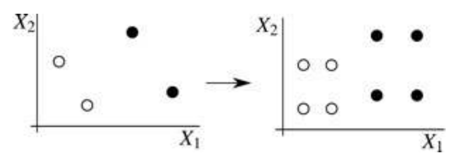
Research Question

How to use such prior knowledge in predictive modeling?

Idea & Results

Data augmentation to reflect the CI relations.

(Idea: Independent ⇨ Shuffled data is equally likely)



- Theory implying the method mitigates over-fitting under correct knowledge of the CG.
- Empirical performance improvement.

Message

Knowledge of causal graphs can be directly used in predictive modeling.

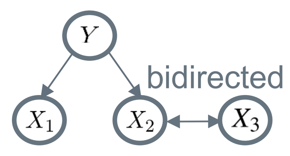
Preliminaries

Acyclic Directed Mixed Graphs (ADMGs) (Richardson, 2003) (Richardson *et al.*, 2017)

Directed acyclic graphs (possibly) with bidirected edges. $\mathcal{G} = ([D], \mathcal{E}, \mathcal{B})$

Used for causal models with latent variables

(semi-Markov models; cf. Latent projection (Tian *et al.*, 2002)).



Topological ADMG Factorization (Tian *et al.*, 2002) (Bhattacharya *et al.*, 2020)

Given a semi-Markov model, $p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})$ holds.

$\mathrm{mp}(j)$: "Markov pillow" of variable Z^j (Generalization of "parents" in ADMGs.)

Problem Setup and Goal

$\mathbf{Z} = (Z^1, \ldots, Z^D) \sim p$: joint data of X and Y.

(each Z^j may be continuous or discrete)

Main Assumption

- $p(\mathbf{Z})$ satisfies the topological ADMG factorization w.r.t. $\mathcal{G}$ (Bhattacharya *et al.*, 2020)

We are given:

- Labeled data $\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p$.
- Estimator $\hat{\mathcal{G}}$ of the underlying ADMG $\mathcal{G}$.

Goal

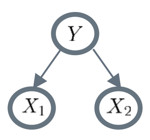
Find a predictor $f : X \mapsto Y$ with small $R(f) = \mathbb{E}[\ell(f, \mathbf{Z})]$.

Key Idea

Idea: Data augmentation to reflect the CI structure.

Example (Trivariate case)

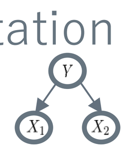
Predict Y from (X_1, X_2), when we know:



(causal graph).

Idea: Data augmentation

The causal graph



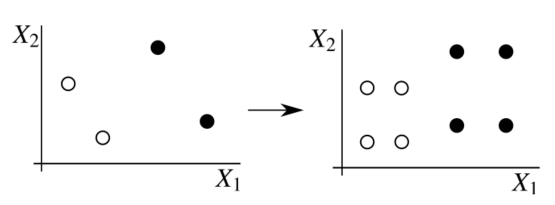
implies $X_1 \perp\!\!\!\perp X_2 \mid Y$.

⇨ Exchange X_1 and X_2 among training samples, stratifying for Y.

Y	X_1	X_2
○	a	c
○	b	d
●	α	γ
●	β	δ

Y	X_1	X_2
○	a	c
○	a	d
○	b	c
○	b	d

Y	X_1	X_2
●	α	γ
●	α	δ
●	β	γ
●	β	δ



Proposed Method Derivation

- Recall topological ADMG factorization: $p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})$.
- Approximate each conditional by kernel-based estimator. Let $K^j : \mathcal{Z}^{\mathrm{mp}(j)} \to \mathbb{R}_{\geq 0}$ and

$$p(\mathbf{Z}) \simeq \prod_{j=1}^{D} \hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) := \frac{\sum_{i=1}^{n} \delta_{Z_i^j}(Z^j) K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}$$

Empirical conditional density

- Plug-in risk estimator

$$\hat{R}_{\mathrm{aug}}(f) = \int_{\mathcal{Z}} \ell(f, \mathbf{Z}) \prod_{j=1}^{D} \hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) \mathrm{d}\mathbf{Z} = \sum_{\mathbf{i} \in [n]^D} \hat{w}_{\mathbf{i}} \cdot \ell(f, \mathbf{Z}_{\mathbf{i}})$$

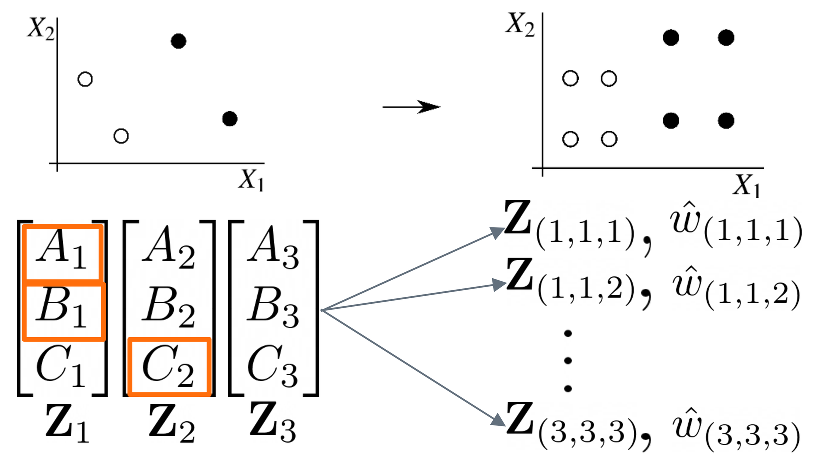
Augmented data + instance weights

Proposed Method

- The plug-in risk estimator can be rewritten as:

$$\hat{R}_{\mathrm{aug}}(f) = \sum_{\mathbf{i} \in [n]^D} \hat{w}_{\mathbf{i}} \cdot \ell(f, \mathbf{Z}_{\mathbf{i}}) \quad \text{where} \quad \mathbf{Z}_{\mathbf{i}} = \begin{bmatrix} Z_{i_1}^1 \\ \vdots \\ Z_{i_D}^D \end{bmatrix}, \quad \hat{w}_{\mathbf{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}_{\mathbf{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}_{\mathbf{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}$$

- This can be computed by data augmentation:



Generate n^D data points (with replacement) with instance weights.

$\mathbf{Z}_{(1,1,2)} = \begin{bmatrix} A_1 \\ B_1 \\ C_2 \end{bmatrix}$

$$\hat{w}_{\mathbf{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}_{\mathbf{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}_{\mathbf{i}_{1:j-1}}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}$$

Theoretical Analysis

Q. How does the proposed method help, statistically?

Setup & Key Assumptions

- True CG does exist, and we have access to it: $\hat{\mathcal{G}} = \mathcal{G}$.
- The underlying densities and the kernel functions satisfy sufficient smoothness and boundedness conditions.

Theorem (Excess Risk Bound; informal) $\hat{f} \in \arg\min_{f \in \mathcal{F}}\{\hat{R}_{\mathrm{aug}}(f)\}$, $f^* \in \arg\min_{f \in \mathcal{F}}\{R(f)\}$

$$R(\hat{f}) - R(f^*) \leq \underbrace{C_1 R_{\mathbf{H}} + C_p}_{\text{Kernel Bias}} + \underbrace{C_2 R_K + C_3 R_{\mathcal{F},K}}_{\text{Complexity terms}} + \underbrace{C_4 \sqrt{\frac{\log(4D/\delta)}{2n}}}_{\text{Uncertainty}}$$

w/ high probability.

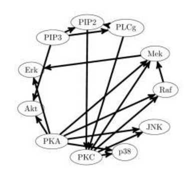
- The complexity terms have a better sample-size dependency than the usual Rademacher complexity, implying mitigated overfitting. (Intuition: Synthesized data ⇨ Reduced possibility of overfitting.)
- But the bias due to the kernel approximation is introduced.

Experiments

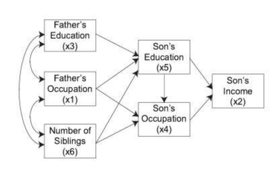
Data and Procedure

▷ 6 data sets from UCI repository (Dua *et al.*, 2017).

▷ 2 data (Sachs and GSS) have reference CGs.



Sachs data (Biology) (Sachs *et al.*, 2005)



GSS data (Sociology) (Shimizu *et al.*, 2011) (Duncan *et al.*, 1972)

NAME	#VAR	#OBS
Sachs	11	853
GSS	6	1380
Boston Housing	14	506
Auto MPG	7	392
White Wine	12	4898
Red Wine	12	1599

▷ DirectLiNGAM was applied to the other data sets to obtain the CGs. (Shimizu *et al.*, 2011)

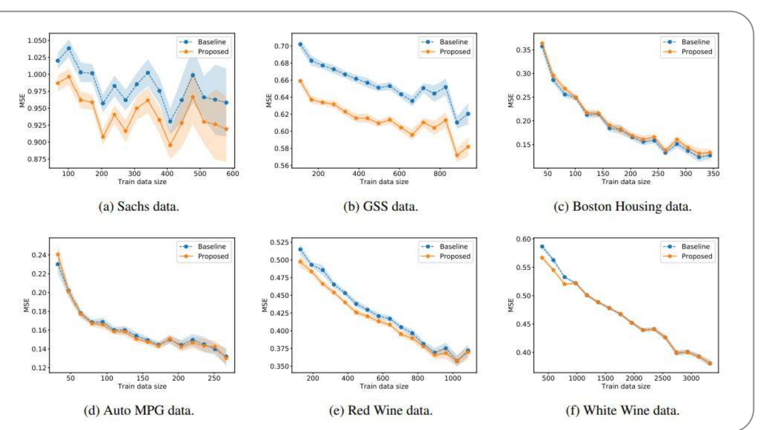
Models and Comparison

Predictor class: gradient boosted regression trees.

Baseline: plain supervised learning $\hat{f} \in \arg\min_{f \in \mathcal{F}}\{\hat{R}_{\mathrm{emp}}(f) + \Omega(f)\}$.

Experiment Results

Improved performance in the small-data regime especially when a CG from domain knowledge is available.



(a) Sachs data. (b) GSS data. (c) Boston Housing data. (d) Auto MPG data. (e) Red Wine data. (f) White Wine data.

References

J. Pearl, *Causality: Models, Reasoning and Inference*, Second. Cambridge, U.K. ; New York: Cambridge University Press, 2009.

K. Sachs *et al.*, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

S. Shimizu *et al.*, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *Journal of Machine Learning Research*, vol. 12, no. 33, pp. 1225–1248, 2011.

O. D. Duncan *et al.*, *Socioeconomic Background and Achievement*, ser. Socioeconomic Background and Achievement. New York: Seminar Press, 1972.

T. Richardson, "Markov properties for acyclic directed mixed graphs," *Scandinavian Journal of Statistics*, vol. 30, no. 1, pp. 145–157, 2003.

T. S. Richardson *et al.*, "Nested Markov properties for acyclic directed mixed graphs," *arXiv:1701.06686 [stat.ME]*, Jan. 2017. arXiv: 1701.06686 [stat.ME].

R. Bhattacharya *et al.*, "Semiparametric inference for causal effects in graphical models with hidden variables," *arXiv:2003.12659 [stat.ML]*, Mar. 2020. arXiv: 2003.12659 [stat.ML].

J. Tian *et al.*, "A general identification condition for causal effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, Menlo Park, CA: AAAI Press/The MIT Press, Aug. 2002, pp. 567–573.

D. Dua *et al.*, *UCI machine learning repository [http://archive.ics.uci.edu/ml]*, Irvine, CA: University of California, School of Information and Computer Science, 2017.

arXiv:2103.00136

37th Conference on Uncertainty in Artificial Intelligence
July 27-30, 2021
Online
uai2021