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

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

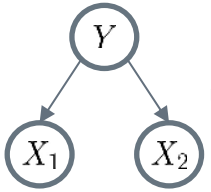
[1]The University of Tokyo [2] RIKEN

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

Causal Graphs (CGs) (Pearl, 2009)

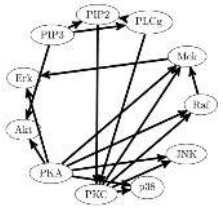
Representation of our knowledge of data generating processes.

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

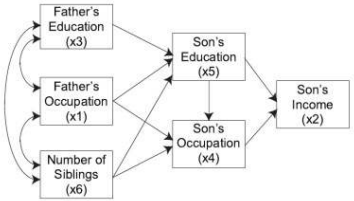


$\Rightarrow X_1 \perp\!\!\!\perp X_2 \mid Y$

Examples:



Biology (Sachs *et al.*, 2005)



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

Research Question

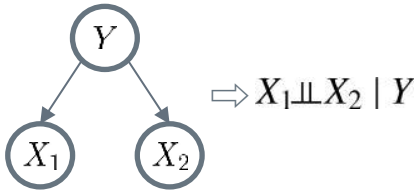
How to use such prior knowledge in predictive modeling?

Background: Causal Graphs

Causal Graphs (CGs) (Pearl, 2009)

Representation of our knowledge of data generating processes.

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

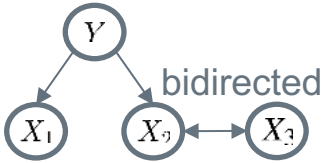


Acyclic Directed Mixed Graphs (ADMGs) (Richardson, 2003) (Richardson *et al.*, 2017)

Directed acyclic graphs (possibly) with bidirected edges.

Used for causal models with latent variables
(semi-Markov models; cf. Latent projection (Tian *et al.*, 2002)).

$$\mathcal{G} = ([D], \mathcal{E}, \mathcal{B})$$



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:

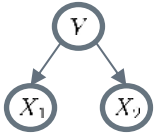
- Labeled data $\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$.
- Estimator $\hat{\mathcal{G}}$ of the underlying ADMG $\mathcal{G}$.

Goal

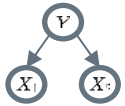
Find a predictor $f : X \mapsto Y$ with small $R(f) = \mathbb{E}[\ell(f, \mathbf{Z})]$.

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

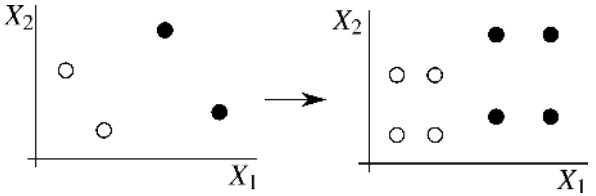
The causal graph  implies $X_1 \perp\!\!\!\perp X_2 \mid Y$.

⇨ 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



Q. How about general graphs?

Proposed Method for general ADMGs

- 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

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

Augmented data + instance weights

Proposed Method for general ADMGs

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

$$\hat{R}_{\mathrm{aug}}(f) = \sum_{\boldsymbol{i} \in [n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$$

where

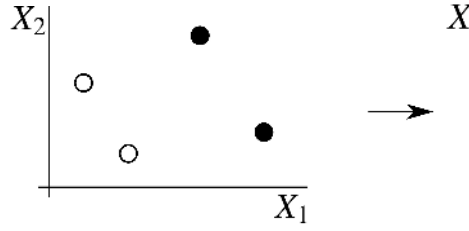
$$\mathbf{Z}_{\mathbf{i}} = \begin{bmatrix} Z^1_{i_1} \\ \vdots \\ Z^D_{i_D} \end{bmatrix} \qquad \hat{w}_{\boldsymbol{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{i})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{k})}$$

$$\boldsymbol{i} = (i_1, \ldots, i_D)$$

- This can be computed by data augmentation:

$$\begin{bmatrix} A_1 \\ B_1 \\ C_1 \end{bmatrix} \begin{bmatrix} A_2 \\ B_2 \\ C_2 \end{bmatrix} \begin{bmatrix} A_3 \\ B_3 \\ C_3 \end{bmatrix} \rightarrow \mathbf{Z}_{(1,1,2)} = \begin{bmatrix} A_1 \\ B_1 \\ C_2 \end{bmatrix}$$

$\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \mathbf{Z}_3$



$$\begin{bmatrix} A_1 \\ B_1 \\ C_1 \end{bmatrix} \begin{bmatrix} A_2 \\ B_2 \\ C_2 \end{bmatrix} \begin{bmatrix} A_3 \\ B_3 \\ C_3 \end{bmatrix} \rightarrow \mathbf{Z}_{(1,1,1)}, \hat{w}_{(1,1,1)}; \quad \mathbf{Z}_{(1,1,2)}, \hat{w}_{(1,1,2)}; \quad \vdots \quad \mathbf{Z}_{(3,3,3)}, \hat{w}_{(3,3,3)}$$

$\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \mathbf{Z}_3$

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)

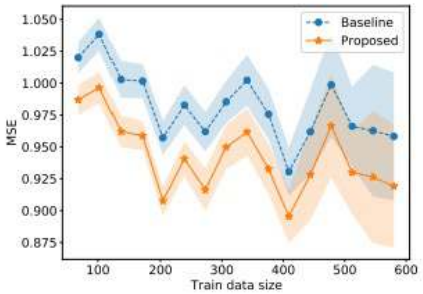
$\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\mathrm{aug}}(f)\}$, $f^* \in \arg\min_{f \in \mathcal{F}} \{R(f)\}$

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

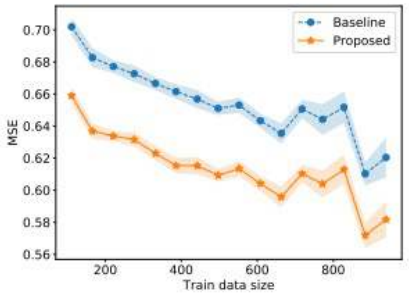
w/ high probability.

- 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.

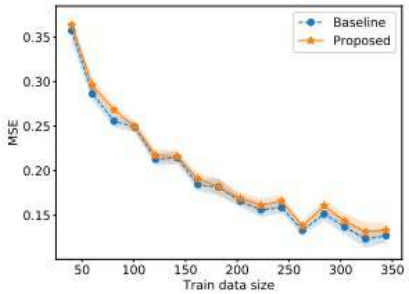
Experimental Results



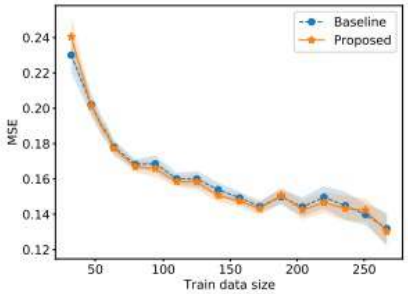
(a) Sachs data.



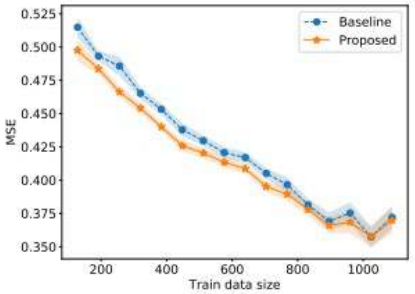
(b) GSS data.



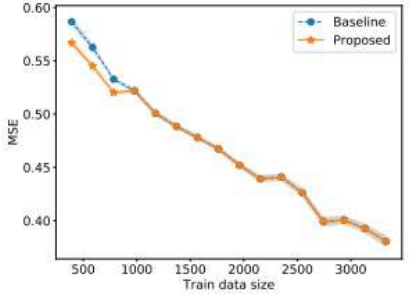
(c) Boston Housing data.



(d) Auto MPG data.



(e) Red Wine data.

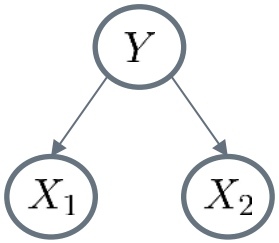


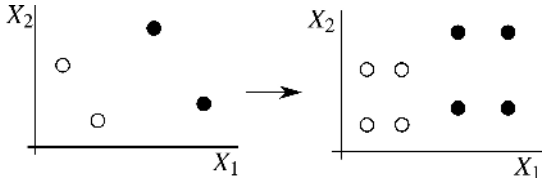
(f) White Wine data.

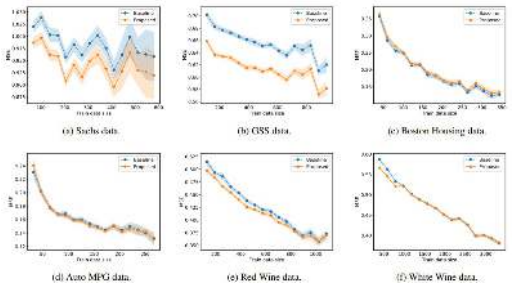
- Improved performance in the small-data regime.

Conclusion

- Proposed a data augmentation method to use the causal graphical prior knowledge in predictive modeling.
- Proposed method suppresses overfitting by inflating the data points but extra bias is introduced by the kernel approximation.
- Experimentally, the benefit may be worth the extra complexity and bias in small-data regime when domain knowledge is available.







Appendix

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.

References

M. A. Little *et al.*, "Causal bootstrapping," *arXiv:1910.09648 [cs.LG]*, Jan. 2020. arXiv: 1910.09648 [cs.LG].

T. Kyono *et al.*, "Improving model robustness using causal knowledge," *arXiv:1911.12441 [cs.LG]*, Nov. 2019. arXiv: 1911.12441 [cs.LG].

T. Kyono *et al.*, "CASTLE: Regularization via auxiliary causal graph discovery," in *Advances in Neural Information Processing Systems 33*, 2020.

S. Magliacane *et al.*, "Domain adaptation by using causal inference to predict invariant conditional distributions," in *Advances in Neural Information Processing Systems 31*, S. Bengio *et al.*, Eds., Curran Associates, Inc., 2018, pp. 10 846–10 856.

M. Rojas-Carulla *et al.*, "Invariant models for causal transfer learning," *Journal of Machine Learning Research*, vol. 19, no. 36, pp. 1–34, 2018.

I. Tsamardinos *et al.*, "Towards principled feature selection: Relevancy, filters and wrappers," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Morgan Kaufmann Publishers, 2003.

K. Yu *et al.*, "Causality-based feature selection: Methods and evaluations," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–36, Sep. 2020.