Distributionally Robust Skeleton Learning of Discrete Bayesian Networks

Yeshu Li [1] Brian D. Ziebart [1]

[1]Department of Computer Science
University of Illinois at Chicago

Nov. 2023

Bayesian Network Skeleton Learning

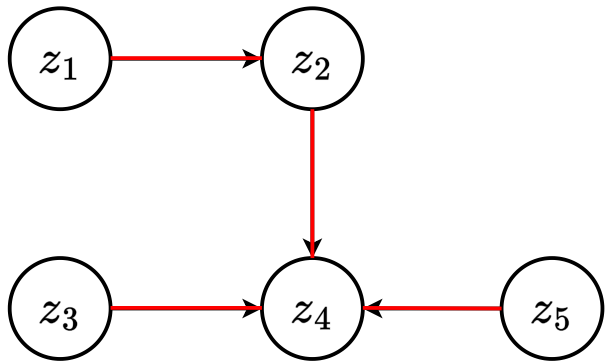
Discrete Bayesian Networks

Given n discrete random variables and a directed acyclic graph $\mathcal{G}$ with parent set $\mathbf{Pa}_i$ for node i associated with the i-th random variable. A Bayesian network is a distribution factorized according to $\mathcal{G}$:

$$\mathbb{P}(\boldsymbol{X}) = \mathbb{P}(X_1, X_2, \ldots, X_n) \triangleq \prod_{i=1}^{n} \mathbb{P}(X_i | \mathbf{Pa}_i).$$

Skeleton Learning / Structure Learning

$$\begin{matrix} z_1 & z_2 & z_3 & z_4 & z_5 \end{matrix}$$
$$\begin{pmatrix} 1 & 3 & 2 & 5 & 6 \\ 2 & 4 & 2 & 1 & 1 \\ & & \cdots & & \\ 1 & 3 & 3 & 1 & 5 \end{pmatrix} \longrightarrow$$



Existing Methods

- **Score-based**: search, dynamic programming, integer linear programming, continuous optimization
- **Constraint-based**: conditional independence tests
- **Hybrid**: leveraging constraint-based methods to restrict the search space of a score-based method

Motivation

Existing methods are either

- relying on **strong assumptions** (faithfulness)
- **computationally expensive** (exponential in the number of nodes)
- **non-robust** (sensitive to data corruption that has large impact on statistical tests)

Baseline: a Surrogate Approach

Given encodings $\mathcal{E}$ and a linear structural equation model (SEM) for each X_r:

$$\mathcal{E}(X_r) = \boldsymbol{W}^{*\mathsf{T}}\mathcal{E}(\boldsymbol{X}_{\bar{r}}) + \boldsymbol{e}.$$

With true distribution $\mathbb{P}$:

$$\boldsymbol{W}^* \in \arg\inf_{\boldsymbol{W}} \frac{1}{2}\mathbb{E}_{\mathbb{P}}\|\mathcal{E}(X_r) - \boldsymbol{W}^{\mathsf{T}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2 \quad \text{s.t.} \quad \boldsymbol{W}_i = \boldsymbol{0} \quad \forall i \in \mathbf{Co}_r.$$

With empirical distribution $\tilde{\mathbb{P}}_m$, a natural regularized empirical risk minimization (ERM) objective for finite samples to estimate $\boldsymbol{W}^*$ is

$$\tilde{\boldsymbol{W}} \in \arg\inf_{\boldsymbol{W}} \tilde{L}(\boldsymbol{W}) := \frac{1}{2}\mathbb{E}_{\tilde{\mathbb{P}}_m}\|\mathcal{E}(X_r) - \boldsymbol{W}^{\mathsf{T}}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2 + \tilde{\lambda}\|\boldsymbol{W}\|_{B,2,1}.$$

Note that the true model does not have to follow a linear SEM form. This is merely for deriving theoretical results.

Distributionally Robust Skeleton Learning

Learning the Bayesian network skeleton with **distribution robust optimization (DRO)**:

$$\hat{\boldsymbol{W}} \in \arg\inf_{\boldsymbol{W}} \sup_{\mathbb{Q}\in\mathcal{A}} \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\|\mathcal{E}(X_r) - \boldsymbol{W}^{\intercal}\mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2.$$

$\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$ is an **ambiguity set** typically defined as

$$\mathcal{A}_{\varepsilon}^{\mathsf{div}}(\tilde{\mathbb{P}}) := \{\mathbb{Q} \in \mathcal{P}(\mathcal{X}) : \mathsf{div}(\mathbb{Q}, \tilde{\mathbb{P}}) \leq \varepsilon\}$$

- discrepancy measure $\mathsf{div}(\cdot, \cdot)$
- ambiguity radius ε

Wasserstein DRO: Formulation

Let the ambiguity set be defined by the Wasserstein distances:

$$W_p(\mathbb{P},\mathbb{Q}) := \inf_{\Pi\in\mathcal{M}(\mathcal{X}^2)} \left\{ \left[\int_{\mathcal{X}^2} c^p(\boldsymbol{x},\boldsymbol{x}')\Pi(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{x}') \right]^{\frac{1}{p}} : \right.$$
$$\left. \Pi(\mathrm{d}\boldsymbol{x},\mathcal{X}) = \mathbb{P}(\mathrm{d}\boldsymbol{x}), \Pi(\mathcal{X},\mathrm{d}\boldsymbol{x}') = \mathbb{Q}(\mathrm{d}\boldsymbol{x}') \right\}.$$

The dual problem of the primal DRO problem can be written as

$$\inf_{\boldsymbol{W},\gamma\geq 0} \gamma\varepsilon + \frac{1}{m}\sum_{i=1}^{m} \sup_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{2}\|\mathcal{E}(x_r) - \boldsymbol{W}^\intercal\mathcal{E}(\boldsymbol{x}_{\bar{r}})\|_2^2 - \gamma\|\mathcal{E}(\boldsymbol{x}) - \mathcal{E}(\boldsymbol{x}^{(i)})\|.$$

Wasserstein DRO: Algorithm

Proposition: NP-hardness

The inner supremum problem of the Wasserstein DRO is NP-hard.

Algorithm 1 Greedy Algorithm for the Wasserstein Worst-case Risk

Input: $\boldsymbol{W}$, γ, $\boldsymbol{x}^{(i)}$
Output: a solution $\hat{\boldsymbol{x}}$
for all $(j, x_j^t) \in [n] \times \mathcal{C}_j$ **do**
  Get a random permutation $\boldsymbol{\pi}$ over $[n]$ with $\pi_1 = j$
  for $k := 2$ **to** n **do**
    $x^t_{\pi_j} \leftarrow \arg\sup_{x^t_{\pi_k}} \ell_{\boldsymbol{W}}(\boldsymbol{x}^t_{\boldsymbol{\pi}_{[k]}}) - \gamma \|\mathcal{E}(\boldsymbol{x}^t_{\boldsymbol{\pi}_{[k]}}) - \mathcal{E}(\boldsymbol{x}^{(i)}_{\boldsymbol{\pi}_{[k]}})\|$
  end for
  if $\boldsymbol{x}^t$ yields a greater objective than $\hat{\boldsymbol{x}}$ **then**
    $\hat{\boldsymbol{x}} \leftarrow \boldsymbol{x}^t$
  end if
end for

Wasserstein DRO: Equivalence to the Baseline

Proposition: Equivalence to Regularized ERM

Let $\ddot{\boldsymbol{W}} := [\boldsymbol{W}; -\boldsymbol{I}_{\rho_r}]^\intercal \in \mathbb{R}^{\rho_{[n]} \times \rho_r}$ with $\boldsymbol{W}_r = -\boldsymbol{I}_{\rho_r}$. If $\gamma \geq \rho_{[n]} \|\ddot{\boldsymbol{W}}\|_F^2$, the Wasserstein DRO problem is equivalent to

$$\inf_{\boldsymbol{W}} \mathbb{E}_{\tilde{\mathbb{P}}_m} \frac{1}{2} \|\mathcal{E}(X_r) - \boldsymbol{W}^\intercal \mathcal{E}(\boldsymbol{X}_{\bar{r}})\|_2^2 + \varepsilon \rho_{[n]} \|\ddot{\boldsymbol{W}}\|_F^2,$$

which subsumes a linear regression approach regularized by the Frobenius norm as a special case.

Wasserstein DRO: Assumptions

Assumptions:

Assumption 1

For the error vector, $\|\boldsymbol{e}\|_\infty \leq \sigma$ and $\|\mathbb{E}_{\mathbb{P}}[|\boldsymbol{e}|]\|_\infty \leq \mu$.

Assumption 2

For each node r, the minimum norm of the true weight matrix $\boldsymbol{W}^*$ for neighbor nodes is lower bounded: $\min_{i \in \mathbf{Ne}_r} \|\boldsymbol{W}_i\|_F \geq \beta > 0$.

Assumption 3

For each node r, $\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r} \succ 0$, or equivalently, $\Lambda_{\min}(\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}) \geq \Lambda > 0$ where $\Lambda_{\min}(\cdot)$ denotes the minimum eigenvalue.

Assumption 4

For each node r, $\|\boldsymbol{H}_{\mathcal{S}_r^c\mathcal{S}_r}\boldsymbol{H}_{\mathcal{S}_r\mathcal{S}_r}^{-1}\|_{B,1,\infty} \leq 1 - \alpha$ for some $0 < \alpha \leq 1$.

Wasserstein DRO: Main Results

Theorem 1

Assume $\|\boldsymbol{W}^*\|_{B,2,1} \leq \bar{B}$ holds for some $\bar{B} > 0$ associated with an optimal Lagrange multiplier $\lambda_B^* > 0$. Let $\varepsilon = \varepsilon_0/m$ where m is the number of samples. Under Assumptions 1-4, if the number of samples satisfies

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{\max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}\right),$$

where C depends on α, Λ, and if $\frac{32\mu\rho_{\max}}{\alpha} < \lambda_B^* < \frac{\beta}{(\alpha/(4-2\alpha)+2)\rho_{\max}\sqrt{\rho_{[n]}}}\sqrt{\frac{\Lambda}{4}}$, then any $\delta \in (0, 1]$, $r \in [n]$, with probability at least $1 - \delta$:

(a) The optimal estimator $\hat{\boldsymbol{W}}$ is unique.

(b) All the non-neighbor nodes are excluded: $\mathbf{Co}_r \subseteq \hat{\mathbf{Co}}_r$.

(c) All the neighbor nodes are identified: $\mathbf{Ne}_r \subseteq \hat{\mathbf{Ne}}_r$.

(d) The true skeleton is successfully reconstructed: $\mathcal{G}_{\text{skel}} = \hat{\mathcal{G}}_{\text{skel}}$.

Kullback-Leibler DRO: Formulation

Let the ambiguity set be defined by the KL divergence:

$$D(\mathbb{Q} \parallel \mathbb{P}) := \int_{\mathcal{X}} \ln \frac{\mathbb{Q}(\mathrm{d}\boldsymbol{x})}{\mathbb{P}(\mathrm{d}\boldsymbol{x})} \mathbb{Q}(\mathrm{d}\boldsymbol{x}).$$

The dual DRO problem becomes

$$\inf_{\boldsymbol{W}, \gamma > 0} \gamma \ln \left[\frac{1}{m} \sum_{i \in [m]} e^{\frac{1}{2} \| \mathcal{E}(x_r^{(i)}) - \boldsymbol{W}^\intercal \mathcal{E}(\boldsymbol{x}_{\bar{r}}^{(i)}) \|_2^2 / \gamma} \right] + \gamma \varepsilon,$$

a convex minimization problem.

Kullback-Leibler DRO: Main Results

Theorem 2

Suppose that $\hat{\boldsymbol{W}}$ is a DRO risk minimizer with the KL divergence and an ambiguity radius $\varepsilon = \varepsilon_0/m$. Given the same definitions of $(\mathcal{G}, \mathbb{P})$, $\mathcal{G}_{\text{skel}}$, $\bar{B}$, λ_B^*, m in Theorem 1. Under Assumptions 1-4, if the number of samples satisfies

$$m = \mathcal{O}\left(\frac{C(\varepsilon_0 + \log(n/\delta) + \log \rho_{[n]})\sigma^2 \rho_{\max}^4 \rho_{[n]}^3}{\min(\mu^2, 1)}\right).$$

where C depends on α, Λ while independent of n, and if the Lagrange multiplier satisfies the same condition as in Theorem 1, then for any $\delta \in (0, 1]$, $r \in [n]$, with probability at least $1 - \delta$, the properties (a)-(d) in Theorem 1 hold.

Experiments: Setup

Contamination models:

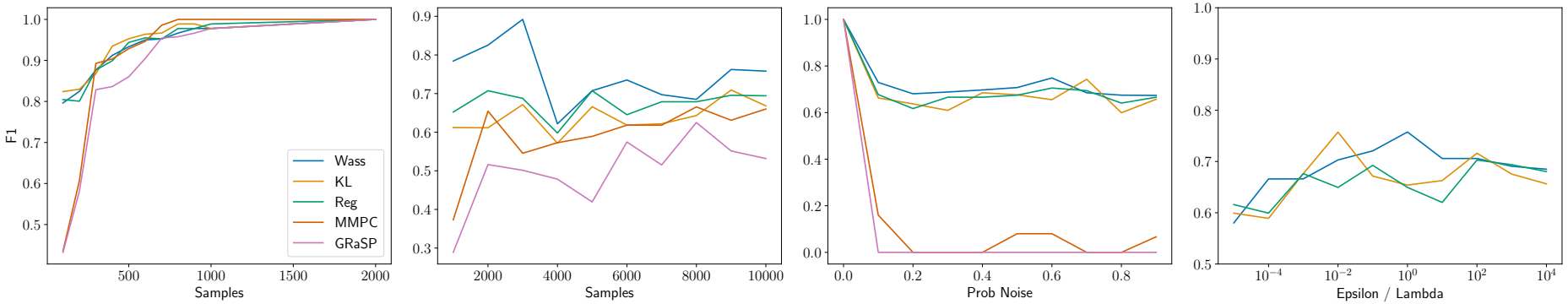
- Noise-free model
- Huber's contamination model
- Independent failure model

Datasets:

- Bayesian Network Repository:
  `https://www.bnlearn.com/bnrepository/`
- Bayesian Network Portfolio by Malone et al.:
  `http://bnportfolio.cs.helsinki.fi/`

Experiments: Results

Dataset	n	m	Noise	ζ	Wass	KL	Reg	MMPC	GRASP	Wass+HC	KL+HC	Reg+HC	MMPC+HC	GRASP+HC	HC
asia	8	1000	Noisefree	0	0.7800†	0.7285†	0.7897†	**0.9067**	**0.8167**	0.5123	0.6367	0.5743	**0.6667**	**0.6583**	0.6550
asia	8	1000	Huber	0.2	**0.7333**†	0.7124†	**0.7297**†	0.5468	0.6570	**0.3943**	**0.3724**	0.3487	0.2907	0.3664	0.2183
asia	8	1000	Independent	0.2	**0.6933**	0.6797	**0.6868**	0.6359	0.3632†	**0.2676**	**0.2632**	0.2581	0.2469	0.1794	0.2443
cancer	5	1000	Noisefree	0	**1.0000**†	**1.0000**†	**1.0000**†	0.6133	0.6133	**0.2800**	**0.2800**	**0.2800**	**0.2800**	**0.2800**	**0.2800**
cancer	5	1000	Huber	0.5	**0.9156**†	0.8933†	**0.9092**†	0.6133	0.5357	**0.4333**	0.3833	**0.4143**	0.2589	0.2714	0.2589
cancer	5	1000	Independent	0.2	**0.9048**†	**0.9029**†	0.8992†	0.0000	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
earthquake	5	1000	Noisefree	0	0.8447†	0.9333†	**0.9778**	**1.0000**	**0.9778**	0.2000	**0.2500**	**0.2500**	**0.2500**	**0.2500**	0.2278†
earthquake	5	1000	Huber	0.2	**0.7509**†	**0.7509**†	**0.7509**†	0.5978	0.6583†	**0.4618**	**0.4618**	**0.4618**	0.3860	0.4547	0.3860
earthquake	5	1000	Independent	0.2	**0.6786**†	**0.6350**†	**0.6350**†	0.0000	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
sachs	11	1000	Noisefree	0	0.8357†	**0.8402**†	0.8374†	**0.9697**	0.7678†	0.4310†	0.4535†	0.4641†	**0.5935**	0.4112†	**0.5873**
sachs	11	1000	Huber	0.2	0.7765	**0.8064**	**0.7893**	0.7498	0.5663†	**0.5194**	0.4815	0.4520	0.4736	0.2380	**0.5028**
sachs	11	1000	Independent	0.5	**0.5268**†	**0.5208**†	0.5172†	0.0000	0.0000	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
survey	6	1000	Noisefree	0	**0.6596**	**0.6545**	0.6506	0.6533	0.1714†	**0.1789**	**0.1789**	**0.1789**	**0.1789**	0.0571	**0.1789**
survey	6	1000	Huber	0.2	**0.7303**†	0.6778†	**0.7095**†	0.5396	0.3810	**0.1444**	**0.1444**	**0.1444**	**0.1444**	**0.1516**	**0.1444**
survey	6	1000	Independent	0.2	**0.6311**†	**0.6705**†	0.6220†	0.2032	0.0000†	**0.1071**	**0.1071**	**0.1143**	**0.1071**	0.0000	**0.1071**
alarm	37	1000	Noisefree	0	0.4750†	0.7863†	**0.8042**†	**0.8530**	0.6824†	0.3483†	0.4949†	0.4470†	**0.5635**	**0.4976**	0.4494†
alarm	37	1000	Huber	0.2	0.1432†	0.1619†	**0.6571**†	**0.5486**	0.1945†	0.2192	0.1680†	**0.3148**	**0.2774**	0.2092†	0.2582
alarm	37	1000	Independent	0.2	0.1419†	0.1448†	**0.5458**†	**0.4309**	0.2830†	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
voting	17	216	Noisefree	0	N/A	N/A	N/A	N/A	N/A	**−2451.8631**	**−2453.2737**	−2453.4091	−2475.5799	−2482.3835	−2456.1489
voting	17	216	Huber	0.2	N/A	N/A	N/A	N/A	N/A	**−4418.9731**	**−4418.9731**	−4487.4544	−4450.3941	−4445.0175	**−4418.9731**
voting	17	216	Independent	0.2	N/A	N/A	N/A	N/A	N/A	**−4453.8298**	**−4453.8298**	−4522.5521	−4465.1076	−4473.8612	**−4453.8298**
backache	32	90	Noisefree	0	N/A	N/A	N/A	N/A	N/A	−1729.8364	−1726.8465	**−1710.7248**	−1719.5002	**−1713.7583**	−1729.7991
backache	32	90	Huber	0.2	N/A	N/A	N/A	N/A	N/A	**−3186.5001**	**−3186.5001**	**−3186.5001**	**−3186.5001**	**−3186.5001**	**−3186.5001**
backache	32	90	Independent	0.2	N/A	N/A	N/A	N/A	N/A	**−2800.9386**	**−2800.9386**	**−2800.9386**	**−2800.9386**	**−2800.9386**	**−2800.9386**



Conclusion

We propose a Bayesian network skeleton learning method with

- robustness
- polynomial time complexity
- polynomial sample complexity
- logarithmic sample complexity for bounded-degree graphs
- competitive empirical performance

Future Work

- Efficient implementations and algorithms
- Evaluation results on a larger scale
- Identifiability
- Tighter bounds
- End-to-end learning with deep neural networks

Thank you!