Convex Mixed-Integer Programming for Causal Additive Models with Optimization and Statistical Guarantees

Xiaozhu Zhang*, Nir Keret†, Ali Shojaie*†, Armeen Taeb*

Department of Statistics, University of Washington*; Department of Biostatistics, University of Washington†

Goal

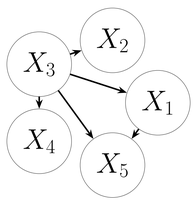
Causal discovery (DAG) for **non-linear** models:

- with **efficiency**
- with **optimality** guarantees
- with **statistical** guarantees
- without equal-variance assumption

Weaknesses of existing methods:

- CAM[1], NoTears[2]: no optimality guarantees
- NPVAR[3]: equal-variance assumption

Model: additive SEM with Gaussian errors



$\iff$

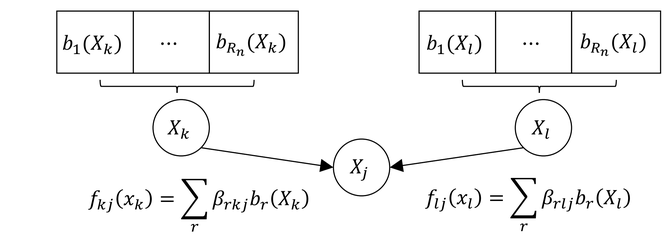
For $j = 1, \dots, p$,

$$X_j = \sum_{k \in \mathrm{pa}(j)} f^\star_{kj}(X_k) + \epsilon_j,$$

$$\epsilon_j \sim \mathcal{N}(0, \sigma_j^{\star 2}), \quad \sigma_j^\star > 0,$$

$$\mathbb{E}[f^\star_{kj}(X_k)] = 0, \ \forall j, k$$

- **Faithfulness**: $f^\star_{kj} \not\equiv 0$ if and only if $k \in \mathrm{pa}(j)$
- **Identifiability of DAG**: when $f^\star_{kj}$ are non-linear and in C^3
- **Using basis functions**: $f_{kj}(x_k) = \sum_{r=1}^{R_n} \beta_{rkj} b_r(X_k)$



$$Z_j = \begin{bmatrix} b_1(X_1) & \cdots & b_1(X_p) & \cdots & b_{R_n}(X_1) & \cdots & b_{R_n}(X_p) \end{bmatrix} \in \mathbb{R}^{n \times pR_n}$$

$$\beta_j = \begin{bmatrix} \beta_{11j} & \cdots & \beta_{1pj} & \cdots & \beta_{R_n 1j} & \cdots & \beta_{R_n pj} \end{bmatrix} \in \mathbb{R}^{pR_n}$$

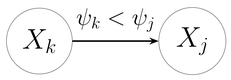
- Group sparsity: If $k \not\to j$, then $\beta_{rkj} = 0, \ \forall r$
- Negative log-likelihood:

$$\ell_n(\beta, \sigma) = \sum_{j=1}^p \log \sigma_j^2 + \sum_{j=1}^p \frac{\|X_j - Z_j \beta_j\|_n^2}{\sigma_j^2}$$

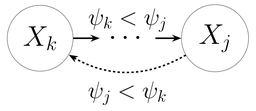
A mixed-integer programming (MIP) formulation

- **Edge selection constraint**:
  Edge presence $g_{kj} \in \{0, 1\}$: $k \not\to j \ \Leftrightarrow \ g_{kj} = 0 \ \Leftrightarrow \ \beta_{rkj} = 0, \ \forall r$
- **Acyclicity constraint**:
  Layer value $\psi_j \in [1, p]$: $1 - p + p g_{kj} \le \psi_j - \psi_k$

direct edges yield layer order



cyclicity yields contradiction



- **Regularization**: $\lambda_n \sum_{kj} g_{kj}$ (DAG sparsity); ridge penalty on β

The MIP formulation g: binary β: continuous

$$\min_{\beta, g, \psi} \ \underbrace{\sum_{j=1}^p \log \left[\|X_j - Z_j \beta_j\|_n^2 + \gamma \|\beta_j\|_2^2 \right]}_{F(\beta, g)} + \lambda_n \sum_{k,j} g_{kj},$$

$$\text{s.t.} \quad (1 - g_{kj}) \beta_{rkj} = 0, \quad \forall r, k, j \quad \text{(edge selection)}$$

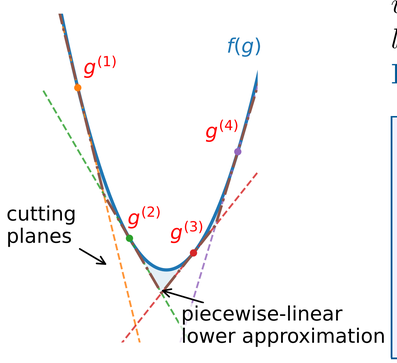
$$1 - p + p g_{kj} \le \psi_j - \psi_k, \quad \forall k, j \quad \text{(acyclicity)}$$

A novel outer approximation (OA) procedure

- Literature: fast OA for linear regression with quadratic MIP[4]
  - First minimize over continuous, and then binary variables
- Derive a novel OA for our non-quadratic MIP w/ the **log term**
  - Prove $f(g)$ can be solved in **closed-form** and it is **convex**:

$$f(g) = \min_\beta F(\beta, g) \ \text{ s.t. edge selection constraint}$$

  - Solve $\min_g f(g)$ s.t. acyclicity constraint via iterative cuts



u: feasible $f(g)$
l: lower approximations
Early stop when $|u - l| \le \tau^{\text{early}}$

A linear MIP in g (scalable)

$$\min_{\theta, g, \psi} \ \theta + \lambda_n \sum_{k,j} g_{kj}$$

$$\text{s.t.} \quad 1 - p + p g_{kj} \le \psi_j - \psi_k, \quad \forall k, j,$$

$$\theta \ge f(g^{(i)}) + \nabla f(g^{(i)})^\top (g - g^{(i)}),$$

$$\forall \text{ cuts } i$$

Theoretical guarantees

Accommodate **non-linearity** and **regularization** simultaneously

Consistency
Under certain regularity conditions, with a proper choice of λ_n, when n is sufficiently large, with probability at least $1 - 6/p$:

1) The optimal **estimated graph** equals the **true graph**;
2) The result holds even with **early stopped estimated graph** when τ^{early} is sufficiently small.

Experiments

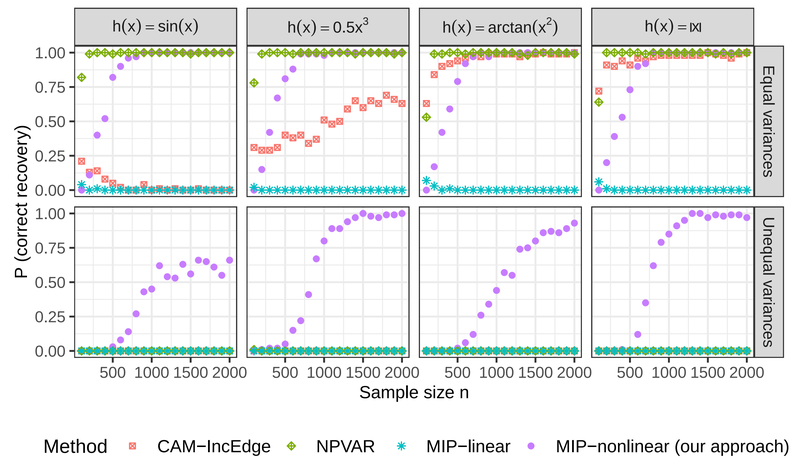


Fig.: Graph recovery performance of baseline methods and our method on a five-variable example. The function $h(x)$ determines the type of non-linearity.

Dataset.$p.s^\star$	MIP OA		CAM		NPVAR	
	Time	*SHD*	*Time*	*SHD*	*Time*	*SHD*
dsep.6.6	0.10	0.3	1.62	2.3	0.14	8.5
asia.8.8	0.22	0.5	2.38	0.0	0.30	12.1
bowling.9.11	0.32	1.1	2.93	1.6	0.42	14
inssmall.16.25	2.18	5.1	7.51	12.5	2.07	36.2
insurance.27.52	56.17	8.3	15.58	16.7	15.3	82

Table: Performance comparison when the SEM is unequal-variance.

References

[1] Bühlmann, P., Peters, J. and Ernest, J. (2014), 'CAM: Causal additive models, high-dimensional order search and penalized regression', The Annals of Statistics 42(6), 2526 – 2556.

[2] Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. Advances in neural information processing systems, 31.

[3] Gao, M., Ding, Y. and Aragam, B. (2020), 'A polynomial-time algorithm for learning nonparametric causal graphs', Advances in Neural Information Processing Systems 33, 11599–11611.

[4] Bertsimas, D., & Van Parys, B. (2020). Sparse high-dimensional regression. The Annals of Statistics, 48(1), 300-323.